

Zero-Shot Learning: An Energy based Approach

Tianxiang Zhao[†], Guiquan Liu^{†*}, Le Wu[‡], Chao Ma[♭], Enhong Chen[†]

[†]Anhui Province Key Laboratory of Big Data Analysis and Application, University of Science and Technology of China

[‡]Hefei University of Technology

[♭]Baidu Talent Intelligence Center

Abstract—Zero-shot learning deals with the problem when the training domain and the test domain have different class sets of image instances. To tackle the problem of some classes in the test data never appeared in the training set, a most popular approach is to map both images and classes in a common space under the embedding based framework. Nevertheless, most embedding based models suffered from the semantic loss problem. Furthermore, the expressive power is limited by representing classes and images as mere points. To tackle these problems, in this paper, we propose an Energy-Based Zero-shot Learning model (EBZL) to encode the association between class attributes and input images for zero-shot learning. EBZL is composed of two parts. The first part is a variational autoencoder that reduces the input dimension of images with representative hidden representations. By feeding the hidden representations as the input of the second part, the second part works as the energy function part based on the deep Boltzmann machine. Specifically, we adapt tradition deep Boltzmann machine to a supervised setting without changing its property as an undirected probabilistic graphic model, which helps to preserve semantic integrity and circumvents semantic loss problem. We further utilize variational inference techniques and mean-field approximation to reduce time complexity in model training process. Finally, extensive experimental results on several real-world datasets clearly show the effectiveness of our proposed method.

Index Terms—zero-shot learning; deep Boltzmann machine

I. INTRODUCTION

With the rapid development of machine learning techniques in recent years, tremendous progress has been made in many machine learning tasks, such as object recognition [10, 20], natural language processing [34], and time-series prediction [9, 22]. However, these approaches confine the system to a closed set of classes and require a large number of supervised training samples to achieve satisfactory classification performance [7]. Well-annotated data is difficult as well as expensive to obtain. Therefore, more and more interests have been paid to the domain of transfer learning, which utilizes supervised information from related tasks [27]. One particular scenario is that there are no available samples for the test class in the training dataset [17]. This situation arises mainly for two reasons. First, samples in the real world often follows a long tail distribution, making it extremely hard and uneconomical to guarantee the presence of all classes, especially those rare ones [21]. Second, the set of classes is ever-growing, and it is impossible to keep the dataset up-to-date [21]. The emergence of new concepts makes it an arduous task to collect a sufficiently large training set.

Zero-Shot Learning [31], abbreviated as ZSL, or zero-shot visual recognition, tries to deal with this problem when the training and testing domains have different class sets. In other words, there is no labeled samples available for the classes in test domain. The goal of ZSL learning is to generalize the classifiers trained on seen classes to these unseen classes. Therefore, the key challenge of ZSL is how to model the correlation among different classes, and transfer knowledge from seen classes to unseen ones. Furthermore, how to build classifiers for unseen classes, and guarantee that these classification models are discriminative enough.

As shown in Fig.1, existing methods for ZSL usually leverage side information, typically in the form of class attributes to depict the association between different categories [18, 30]. The common paradigm of ZSL models follows an embedding-based framework [8, 39, 40]. These embedding based models learn two mapping functions using the available samples from seen classes to transform both the input images and the classes into a common space. Then, the similarity and distance between each sample and each class can be measured. However, by representing each class as a mere point in the common space, these models neglected the intra-class variability. Therefore, the expressive power is limited. What's worse, they also suffer from the semantic loss problem. In fact, the mapping function from image input space to the embedding space can be seen as a feature extractor. In the training process, these embedding based models naturally discarded features that are less-discriminative for the unseen classes. However, these discarded features could be beneficial for the classification of unseen classes, due to the semantic discrepancy between seen and unseen classes.

Considering the limitation of representing each class as a mere point and measure similarity based on distance in these embedding based models, in this paper, we propose a novel Energy-Based Zero-shot Learning model (EBZL) which returns the compatibility between an input image and the attributes of each class. The proposed model is composed of two parts. Concretely, we first use the classical variational autoencoder to reduce the dimension of input images. Then, we design an undirected graph model named s-DBM, which works as the energy function part to measure the correspondence between an input and a class attribute vector. We embed the supervised information into the deep Boltzmann machine without changing its structure. Specifically, to preserve semantic integrity, we adopt undirected links in the

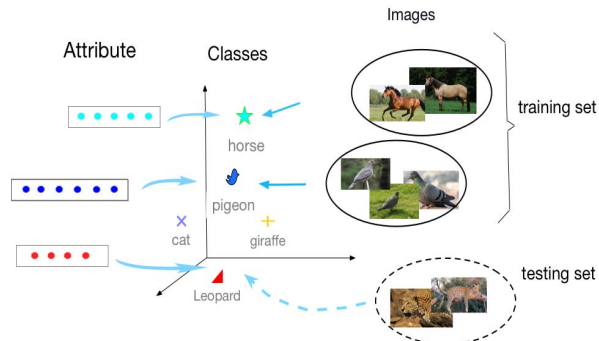


Fig. 1: Illustration of the zero-shot learning problem. In the test stage, we want to learn a classifier for leopard, but we do not have any images that belong to the leopard category in the training data. In ZSL problem, we usually leverage the attributes associated with each class to connect the association between unseen class and other seen classes in the training data.

energy function part. The energy model, which returns the co-occurrence of one point in the input of image space and one in the class space, can represent all sorts of distributions and therefore better model the relationship between images and class attributes [19]. Besides, as each layer of our energy function part has bidirectional links, it can fit the parameters in the top-down manner in addition to conventional bottom-up manner. Therefore, the undirected modeling could better reconstruct inputs from extracted features and prevents the semantic loss problem. In model training process, we utilize the variational inference technique and mean-field approximation to accelerate the optimization step.

We perform experiments on four widely-used datasets, AwA [17], CUB-200 [36], SUN [28], and aPY [6]. The experimental results show an increase of 1.53%, 2.4%, 5.72%, 2.13% over the best state-of-the-art models [11] respectively on each dataset, which validate the effectiveness of our proposed method.

In summary, our main contributions are three-fold:

- We address the ZSL problem using an energy-based model to encode the association between class attribute vectors and input images, which is much more expressive than traditional methods that represent classes and images as mere points.
- We extend traditional DBM to supervised settings for energy function part. The connections in each layer are undirected, which can model the reconstruction process and solve the semantic loss problem. To accelerate the training process, we further utilize variational inference techniques and mean-field approximation to reduce time complexity.
- We conduct extensive experiments on four widely used datasets for ZSL image classification problem, indicating their superiority against state-of-the-art methods.

II. RELATED WORK AND PRELIMINARIES

In this section, we will first go into the details of previous ZSL methods. After that, we will give a concise review on restricted Boltzmann machine, which is closely related to our proposed method.

A. Related Work

The goal of ZSL problem is to recognize the unseen categories without available samples in the training data. We summarize the existing methods on ZSL into three categories: embedding-based methods, basis-based methods and sample synthesis-based methods.

Embedding-based models learn the mapping of both classes and samples in a common embedding space [8, 39, 40]. These methods can be summarized into a general function as follows: $c(x) = \operatorname{argmax}_y \operatorname{sim}(\tau(\mathbf{x}), \psi(\mathbf{a}_y))$, where x is a sample and c is a classifier. sim measures the similarity, and \mathbf{a}_y refers to the attribute vector of class y . Specifically, τ maps the input images into the embedding space, and ψ maps the classes with the side information of class attributes. This kind of models assume that an image should belong to the class whose attributes have least distance to $\tau(\mathbf{x})$ in the embedding space. ESZSL makes modifications to previous methods by adding some regularization terms to enhance the expressive power of the embedding space [31]. However, there are two limitations of this kind of methods. First is the semantic loss problem. τ can be viewed as a feature extraction function. When training, it only learns to preserve the features that are discriminative for seen classes. Therefore, the extracted features may be not suitable for classifying unseen classes, as some information that is valuable for the unseen classes is discarded in the training process. Second is the simple assumption that maps each class as a mere point, which causes quite restricted expression power. Some attempts have been made to address the semantic loss problem. E.g., [14] introduced a recovery loss in the modeling process to ensure that the learned embeddings could also recover the input data. Researchers designed a semantic preserving adversarial embedding network to prevent the semantic loss by introducing an independent visual to semantic space [4]. However, these embedding based methods suffer from the limited expressive power of modeling each sample as a mere point. Different from them, we design an approach by modeling the distribution of data to both avoid the semantic loss problem and enhance the expressive power.

Instead of aligning the image and attribute domain as most embedding-based methods, the basis-based models learn class-specific classifier for ZSL problem [3, 5, 35]. Basis-based methods can roughly be summarized as training classifiers separately for each seen classes, and using them to synthesize classifiers for unseen classes. They commonly utilize attributes as the guiding information. RIS [35] and DAP [17] train independent classifiers for each attribute, taking the mapping from attributes to classes as a given linear transformation. SCZL [3] chooses a subset of seen classes as bases, and use them to align the semantic space(attributes) and model space(classifiers). It assumes that the classifier of one class can

be expressed based on the classifiers of those bases classes in the same way its attributes being expressed by those bases attributes. These methods all have semantic loss problem and the strong attribute-independent assumption. RULE [1] modifies RIS by using attribute vector as a whole, and uses it to guide the synthesis of classifiers for unseen classes. SMS [13] learns a function which takes semantic attributes as the input and outputs the classifier model. However, the class attributes are correlated, measuring the proximity among classes using only using attributes is inaccurate. Besides, the semantic loss problem also exists in most of these basis-based models.

Synthesis-based methods are relatively new [11, 12, 41]. They train a generative model from attributes to images and synthesize pseudo samples for unseen classes. After that, the ZSL problem could be transformed to the classical supervised classification form. SSZL [11] proposes to generate samples for new classes through learning relationship among classes using their attributes. It assumes that if two class are similar in the attribute space, then their samples should also be similar. FZLC [21] synthesizes samples using a mediate space and builds two mapping functions. One mapping function lies between the input and the mediate space, and the other one between the mediate and the attribute space. Despite their relatively high performance, the generation ability is the bottleneck of this kind of methods. Images of real-life species are usually diverse, so synthesizing high-quality pseudo samples based only on the attributes is difficult. Synthesized images may drift from real ones and consequently harm the effectiveness of the trained classifiers. We build an energy-based model to enhance traditional embedding-based methods. To the best of our knowledge, this is the first work that considers exploiting joint distributions of input images and class attributes. By aligning the input image space and the class space using energy function, the proposed model can depict all sorts of distributions, and are much more expressive than traditional ones.

B. Restricted Boltzmann Machine

Restricted Boltzmann Machine (RBM) is an undirected probabilistic graph model under the unsupervised setting. It is originally composed of two-layered network, with a visual layer (input) and a hidden layer (can be seen as embedding layer). When training, it tries to maximize the joint probability between input images and the hidden vectors with the energy function as:

$$\mathbf{E}(\mathbf{x}, \mathbf{h}) = -(\mathbf{v}^T \mathbf{x} + \mathbf{u}^T \mathbf{h} + \mathbf{x}^T \mathbf{W} \mathbf{h}), \quad (1)$$

where \mathbf{v} , \mathbf{u} , \mathbf{W} are parameters. \mathbf{x} is input and \mathbf{h} is the hidden vector. It is commonly used as a feature extractor to find the most representative hidden vector for the input image. Several works have dedicated to expand this model to supervised problem settings. [25, 26] added a SVM layer above the hidden layer, using RBM as the base feature extractor. [24] proposes a unified framework for supervised RBM, which preserves its property as a probabilistic graphical model,

and can be trained jointly. However, it assumes that the top layer follows a distribution in exponential family. This directed bottom-up connection assumption without top-down connections would impair the model's power as an energy function. As the information could not be propagated from the top-down connections, the semantic integrity of the upper layer can not be guaranteed, and would lead to the semantic loss problem.

As deep structure has been proven useful for many visual tasks, to enhance the modeling power of RBM, Deep Boltzmann Machine (DBM) is proposed by stacking traditional RBMs together using multiple hidden layers [?]. The traditional training algorithms for DBM is the MCMC training algorithm. However, as using MCMC to estimate the probability when there are multiple hidden layers is intractable, researchers also designed the variational inference algorithm [15] and an inference net to accelerate the training process [32].

III. THE PROPOSED MODEL

A. Problem Definition

In the problem setting, we are given S seen classes and U unseen classes, along with attributes of these classes as side information, each being a M -dimensional class-attribute vector. In this work, each class y is represented by its attribute vector \mathbf{a}_y , and attributes of all classes form a matrix $\mathbf{A} \in \mathbb{R}^{M \times (S+U)}$, with the j -th column denotes the attribute vector of the j -th class. The training dataset is denoted as $D_{train} = \{(\mathbf{x}_k, \mathbf{a}_{y_k})\}_{k=1}^N$ with N training samples, where \mathbf{x}_k denotes the k -th input image and is the k -th column of the input matrix \mathbf{X} . y_k is its label denoting the class it belongs to. These training samples are all from seen classes.

Before going further, we will first present the key ideas of our proposed energy-based model for the ZSL problem. As discussed above, energy model is suitable for capturing joint probabilistic distributions of vectors from different domains. So, in this paper we will use it to model how much the input image features are related to each vector in the attribute space, which suits the ZSL problem setting quite well. And, our energy-function part is a symmetrical graph model, which circumvents the semantic loss problem as it is computed bidirectionally. Concretely, as shown in Fig. 2, the model is comprised of two parts, a Variational AutoEncoder (VAE) part and a supervised DBM part (s-DBM) that models the supervision information under the DBM model. The VAE part can extract representative features with sufficient semantic information. By reducing the input information with representative features, this step can significantly reduce the complexity in the training of the s-DBM part. The s-DBM part is an extension of the traditional DBM with supervised information in the top layer. Different from previous works in this direction, we build it as an undirected graphical model, which can be used as the energy function.

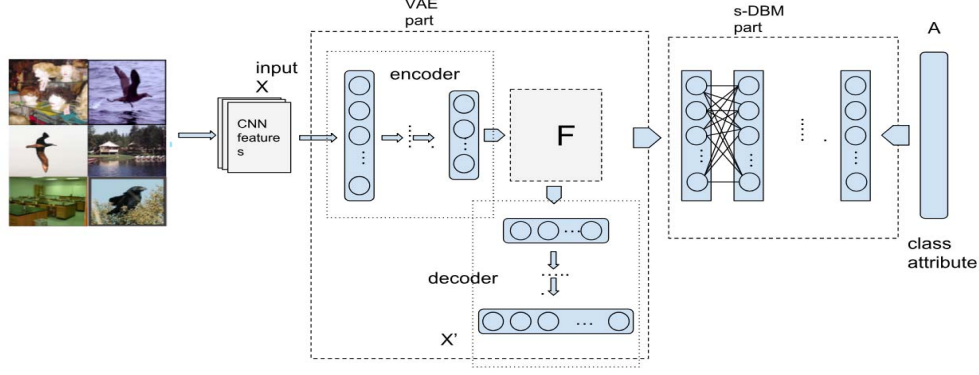


Fig. 2: Overview of our proposed model. We use a VAE to reduce the dimensions of input image features obtained through a VGGnet. Then, we modify DBM to supervised settings (as s-DBM) and use it to model the relationship between input features and attributes. Here, s-DBM part is set as a four-layer structure.

B. VAE Part

VAE is a variant of classical autoencoder models. It assumes that the encoding results (latent variables) follow a prior of Gaussian distribution. Its goal is to use a recognition model to approximate the posterior distribution of the latent variables given the input images, and to train a generative model to recover the original image based on encoding results. As shown in the first part of Fig. 2, it has two components: an encoder and a decoder. The encoder infers the posterior distribution of the latent vector \mathbf{f}_i given the input \mathbf{x}_i , i.e., $p_\theta(\mathbf{f}_i|\mathbf{x}_i)$. The decoder learns to recover the original input \mathbf{x}_i given the encoding result \mathbf{f}_i , i.e., $q_\eta(\mathbf{x}_i|\mathbf{f}_i)$. This model is frequently used in a generative way, but the encoding result \mathbf{f}_i can also be seen as a good representation of original input \mathbf{x}_i , as it preserves low-level features well to reconstruct the original image. We train it by minimizing the traditional VAE loss and the energy value in Equ.11, where the prior $p_0(\mathbf{f})$ is the standard Gaussian distribution. Specifically, the loss function of the VAE part is defined as:

$$\begin{aligned} \mathcal{L} &= -L_{\theta,\eta}(\mathbf{x}_i) + \mathbf{E}(\mathbf{f}_i, \mathbf{a}_i; \phi) \\ &= -E_{p_\theta(\mathbf{f}_i|\mathbf{x}_i)}[\log q_\eta(\mathbf{x}_i|\mathbf{f}_i)] + KL(p_\theta(\mathbf{f}_i|\mathbf{x}_i)||p_0(\mathbf{f})) \\ &\quad + \mathbf{E}(\mathbf{f}_i, \mathbf{a}_i; \phi) \end{aligned} \quad (2)$$

For each image i , similar as many image representation techniques, we first send this image into a convolutional neural network to get the 4096 dimension representation, i.e., \mathbf{x}_i for each image. We use VAE as the feature extractor to learn the feature representation \mathbf{f}_i for input sample \mathbf{x}_i , and send \mathbf{f}_i as input for the second part, i.e., s-DBM, of our model. This VAE part could be seen as a dimension reduction technique for the s-DBM part. Directly using the 4096 dimensional representation would make the s-DBM part have too many parameters to fit. At the same time, too much trivial information makes it hard for s-DBM to converge to a optimal solution. Based on these observations, the VAE part

reduces the dimensions of input images by learning a better representation of the original images.

C. s-DBM Part

As shown in the second part of Fig. 2, s-DBM is an extension of DBM by adding a layer in the top, containing the information of class attributes so that the model can be trained in a supervised way. Note that it is not confined to a four-layer structure, we just stack it on the conventional three-layer DBM to illustrate our idea. Therefore, s-DBM can model the energy $E(\mathbf{f}_i, \mathbf{a}_j)$ for a pair of input images and their associated attribute vectors.

Specifically, considering DBM with supervision information has already been proposed in the past, which assumed a conditional distribution of exponential family with directed connections[24]. Instead, we assume a Boltzmann distribution for $p(\mathbf{a}_j|\mathbf{h}_{i_{tbo}}; \phi)$ with bidirectional connection to ensure that information could be propagated from both top-down and bottom-up. Here, ϕ denotes the parameters and $\mathbf{h}_{i_{tbo}}$ refers to the output of the top-but-one layer in the s-DBM. In this way, we preserve the property of the DBM as an energy function. It can be embedded directly into any DBM structures, and here we give a simple four-layer example as an example for better illustration. In the four layer structure, the energy function can be derived from Equ.1 as:

$$\begin{aligned} \mathbf{E}(\mathbf{f}, \mathbf{h}, \mathbf{a}; \phi) &= -(\mathbf{v}_1^T \mathbf{f} + \mathbf{u}_1^T \mathbf{h}_1 + \mathbf{f}^T \mathbf{W}_1 \mathbf{h}_1) \\ &\quad -(\mathbf{v}_2^T \mathbf{h}_1 + \mathbf{u}_2^T \mathbf{h}_2 + \mathbf{h}_1^T \mathbf{W}_2 \mathbf{h}_2) \\ &\quad -(\mathbf{v}_3^T \mathbf{h}_2 + \mathbf{u}_3^T \mathbf{a} + \mathbf{h}_2^T \mathbf{W}_3 \mathbf{a}), \end{aligned} \quad (3)$$

where \mathbf{h}_i denotes the results of the i -th intermediate layer and $\mathbf{v}_i, \mathbf{u}_i, \mathbf{W}_i$ are detailed model parameters in the parameter set ϕ in corresponding layers. We omit the subscript for \mathbf{f} and \mathbf{a} . In this way, the joint probability can be modeled as:

$$p(\mathbf{f}, \mathbf{h}, \mathbf{a}; \phi) = \exp(-\mathbf{E}(\mathbf{f}, \mathbf{h}, \mathbf{a}; \phi) - \mathbf{A}(\phi)), \quad (4)$$

where $\mathbf{A}(\phi)$ is the log-partition function.

Our aim is to maximize $p(\mathbf{f}_k, \mathbf{a}_{y_k}; \phi)$, which is the expectation of Equ.4 over \mathbf{h} and can be computed using Bayesian rules as in Equ.5. We train the model to give a higher probability to pair \mathbf{f}_k and \mathbf{a}_{y_k} that appeared in the training set, so that we can use this model to find \mathbf{a} which has a high joint probability with given \mathbf{f} as:

$$\begin{aligned} p(\mathbf{f}, \mathbf{a}; \phi) &= \sum_{\mathbf{h}} \mathbf{p}(\mathbf{f}, \mathbf{h}, \mathbf{a}; \phi) \mathbf{p}(\mathbf{h}|\mathbf{f}, \mathbf{a}; \phi), \\ p(\mathbf{h}|\mathbf{f}, \mathbf{a}; \phi) &= \mathbf{p}(\mathbf{f}, \mathbf{h}, \mathbf{a}; \phi) / \sum_{\mathbf{h}} \mathbf{p}(\mathbf{f}, \mathbf{h}, \mathbf{a}; \phi) \end{aligned} \quad (5)$$

To summarize, as shown in the second part of Fig. 2, the s-DBM part contains four layers: input layer F , intermediate layer H_1 , intermediate layer H_2 , and the top layer A . The energy function can be written as in Equ.3, and the probability density is presented in Equ.4. To maximize the value of $p(\mathbf{f}, \mathbf{a}; \phi)$, we first compute its derivative of log-likelihood with respect to parameter vector \mathbf{W}_1 and \mathbf{v}_1 as follows:

$$\begin{aligned} \frac{\partial \log p(\mathbf{f}, \mathbf{a}; \phi)}{\partial \mathbf{W}_1} &= \frac{\sum_{\mathbf{h}} p(\mathbf{f}, \mathbf{h}, \mathbf{a}; \phi) (\mathbf{f} \mathbf{h}_1^T - \frac{\partial \mathbf{A}}{\partial \mathbf{W}_1}) \mathbf{p}(\mathbf{h}|\mathbf{f}, \mathbf{a}; \phi)}{p(\mathbf{f}, \mathbf{a}; \phi)} \\ &= \mathbb{E}_{p_{data}}(\mathbf{f} \mathbf{h}_1^T) - \mathbb{E}_{p_{model}}(\mathbf{f} \mathbf{h}_1^T) \\ \frac{\partial \log p(\mathbf{f}, \mathbf{a}; \phi)}{\partial \mathbf{v}_1} &= \frac{\sum_{\mathbf{h}} p(\mathbf{f}, \mathbf{h}, \mathbf{a}; \phi) (\mathbf{f} - \frac{\partial \mathbf{A}}{\partial \mathbf{v}_1}) \mathbf{p}(\mathbf{h}|\mathbf{f}, \mathbf{a}; \phi)}{p(\mathbf{f}, \mathbf{a}; \phi)} \\ &= \mathbb{E}_{p_{data}}(\mathbf{f}) - \mathbb{E}_{p_{model}}(\mathbf{f}) \end{aligned} \quad (6)$$

where p_{data} denotes the training set data distribution, and p_{model} denotes the model distribution. The derivatives with respect to other parameters take similar forms, involving the expectation of $\mathbf{h}_1 \mathbf{h}_2^T$, $\mathbf{h}_2 \mathbf{a}^T$ and other vectors respectively.

D. Variational Inference

However, it is intractable to compute expectations in Equ.5 exactly, as the computation of $p(\mathbf{h}|\mathbf{f}, \mathbf{a}; \phi)$ is time-consuming. The data-dependent expectation requires the sum over space that is exponential in the number of hidden units, and the model-dependent expectation requires the sum over the space that is exponential in the number of hidden and visible units together. These expectations can be approximated using Gibbs sampling based on Equ.4 and Equ.5, but it still takes a lot of time. So, inspired by [15], we propose a variational inference technique to accelerate that.

In our proposed variational method, the posterior $p(\mathbf{h}|\mathbf{f}, \mathbf{a}; \phi)$ is approximated by $q(\mathbf{h}; \mu)$, where μ is the variational parameter. Following classical variational inference steps, we arrive at following evidence lower bound(ELBO) as:

$$\mathbb{L} = \mathbb{E}_{q(\mathbf{h}; \mu)}(\log p(\mathbf{f}, \mathbf{h}, \mathbf{a}; \phi)) - \mathbb{E}_{q(\mathbf{h}; \mu)} \log q(\mathbf{h}; \mu). \quad (7)$$

We can learn the parameters by maximizing ELBO. Following a naive mean-field approach, we choose a variational

distribution that can be fully factorized into Bernoulli distributions:

$$q(\mathbf{h}; \mu) = \prod_j \prod_k \prod_z \mathbf{q}(\mathbf{h}_{1j}; \mu_{1j}) \mathbf{q}(\mathbf{h}_{2k}; \mu_{2k}) \mathbf{q}(\mathbf{h}_{3z}; \mu_{3z}), \quad (8)$$

where h_{ij} is the j -th dimension of hidden layer h_i , similarly for μ_{ij} . μ also denotes the parameters for corresponding Bernoulli distributions. Now, the ELBO takes a particular simple form:

$$\begin{aligned} \mathbb{L} &= (\mathbf{v}_1^T \mathbf{f} + \mathbf{u}_1^T \mathbf{h}_1 + \mathbf{f}^T \mathbf{W}_1 \mathbf{h}_1) \\ &\quad + (\mathbf{v}_2^T \mathbf{h}_1 + \mathbf{u}_2^T \mathbf{h}_2 + \mathbf{h}_1^T \mathbf{W}_2 \mathbf{h}_2) \\ &\quad + (\mathbf{v}_3^T \mathbf{h}_2 + \mathbf{u}_3^T \mathbf{a} + \mathbf{h}_2^T \mathbf{W}_3 \mathbf{a}) - \mathbf{A}(\phi) + \mathcal{H}(q(\mu)). \end{aligned} \quad (9)$$

With the model parameters ϕ fixed, μ can be explicitly computed by:

$$\begin{aligned} \mu_1 &= \text{sigmoid}((\mathbf{f}^T \mathbf{W}_1)^T + \mathbf{u}_1 + \mathbf{W}_2 \mu_2 + \mathbf{v}_2) \\ \mu_2 &= \text{sigmoid}((\mu_1^T \mathbf{W}_2)^T + \mathbf{u}_2 + \mathbf{W}_3 \mathbf{a} + \mathbf{v}_3). \end{aligned} \quad (10)$$

Now that we can approximate $p(\mathbf{h}|\mathbf{f}, \mathbf{a}; \phi)$ using $q(\mathbf{h}; \mu)$, Equ.6 can be computed, and we are able to update parameters ϕ to train the proposed model.

E. Model Training Part

The VAE part and s-DBM part are trained alternatively. The VAE part is used to reduce the input dimensions and preserve sufficient semantic information, so we follow the traditional training techniques [29]. To train s-DBM part, we have to update parameters to maximize $p(\mathbf{f}, \mathbf{h}, \mathbf{a}; \phi)$ for input and attribute pairs appeared in the training data. However, this procedure is intractable, because the computation of $p(\mathbf{h}|\mathbf{f}, \mathbf{a}; \phi)$ takes time that is exponential in the number of hidden units.

With the variational techniques proposed in Equ.9, now we can choose to use the contrastive divergence(CD) method, which runs short Markov chains to approximate the model expectation. It is divided into two steps. First, we fix the model parameters to approximate the distribution of latent vectors by utilizing Equ.10. Then, we sample latent vectors and update model parameters based on Equ.6. Concretely, for each training example, we first find the value of variational parameter μ based on the current value of \mathbf{f}, \mathbf{a} . To solve Equ.10, we simply cycle through layers and update the variational parameters until they converge. Then, we update parameters ϕ . The updating strategy is quite straightforward. Approximation of data-dependent expectation $\mathbb{E}_{p_{data}}(\cdot)$ is computed by performing a point estimations on sample $\mathbf{f}, \mathbf{h}_1, \mathbf{h}_2, \mathbf{a}$, which is sampled based on $q(\mathbf{h}; \mu)$. In practice, we maintain a set of s sample particles and average over them. The approximation of model-dependent expectation $\mathbb{E}_{p_{model}}(\cdot)$ can be computed similarly, with pairs $\{(\mathbf{x}, \mathbf{a}_y)\}$ to be sampled in advance. This optimization procedure belongs to the class of Robbins-Monro algorithms[38], and guarantees asymptotic convergence. Note that we perform layer-wise pretraining technique before training each layer jointly using variational inference, which is a

widely used trick in training DBM. For better illustration, we list the overall training procedure in Algorithm 1.

When testing, we first run VAE for input \mathbf{x} to get its representation \mathbf{f} . Then, we use s-DBM to predict y by returning the y whose \mathbf{a} has the minimum energy value combined with \mathbf{f} . For each candidate a_j we still need to first approximate the distribution of \mathbf{h} based on Equ.10, just like in training. Then, we compute the energy value of current pair \mathbf{f}, \mathbf{a}_j based on Equ.4. Note that we don't need to use sampling to compute $\mathbf{E}(\mathbf{f}, \mathbf{a}; \phi)$. In practice, we can use variational parameters μ instead, and the value function can now be written as:

$$\begin{aligned} \mathbf{E}(\mathbf{f}, \mathbf{a}; \phi) = & - (\mathbf{v}_1^T \mathbf{f} + \mathbf{u}_1^T \mu_1 + \mathbf{f}^T \mathbf{W}_1 \mu_1) \\ & - (\mathbf{v}_2^T \mu_1 + \mathbf{u}_2^T \mu_2 + \mu_1^T \mathbf{W}_2 \mu_2) \\ & - (\mathbf{v}_3^T \mu_2 + \mathbf{u}_3^T \mathbf{a} + \mu_2^T \mathbf{W}_3 \mathbf{a}). \end{aligned} \quad (11)$$

Algorithm 1 Training procedure of the s-DBM

Input:

input feature matrix \mathbf{F} which is obtained using pretrained VAE and input image matrix \mathbf{X} ;
class-attribute matrix \mathbf{A} ;
class indicator y ;
sequence-sequence similarity matrix $\mathbf{W} = \mathbf{A}^T \mathbf{A}$;
 $\tau_{max} = 10^{10}$, $\rho = 1.1$

Output:

parameters $\mathbf{W}, \mathbf{v}, \mathbf{u}$ for s-DBM model

- 1: **while** not converge **do**
- 2: Pretrain $\mathbf{W}_i, \mathbf{v}_i, \mathbf{u}_i$ for each layer i ;
- 3: Sample \mathbf{f}' and \mathbf{a}' for the approximation of data-independent expectation;
- 4: **while** not converge **do**
- 5: Get input batch of \mathbf{f} and \mathbf{a} ;
- 6: Initialize variational parameter μ_1 and μ_2 ;
- 7: **while** not converge **do**
- 8: Update μ_1 according to Equ.10;
- 9: Update μ_2 according to Equ.10;
- 10: **end while**
- 11: Sample s particles based on $q(\mathbf{h}, \mu)$
- 12: Use these samples to approximate data-dependent expectation in Equ.6;
- 13: Sample s particles based on sampled \mathbf{f}' and \mathbf{a}' ;
- 14: Use these samples to approximate data-independent expectation in Equ.6;
- 15: Update \mathbf{W} for each layer according to Equ.6;
 $\mathbf{W}_i = \mathbf{W}_i + \tau \left(\frac{\partial \log p(\mathbf{f}, \mathbf{a}; \phi)}{\partial \mathbf{W}_i} \right)$;
- 16: Update \mathbf{v} and \mathbf{u} for each layer according to Equ.6;
 $\mathbf{v}_i = \mathbf{v}_i + \tau \left(\frac{\partial \log p(\mathbf{f}, \mathbf{a}; \phi)}{\partial \mathbf{v}_i} \right)$;
 $\mathbf{u}_i = \mathbf{u}_i + \tau \left(\frac{\partial \log p(\mathbf{f}, \mathbf{a}; \phi)}{\partial \mathbf{u}_i} \right)$;
- 17: $\tau = \tau * \rho$;
- 18: Update parameters of VAE following Equ.2;
- 19: Obtain new input feature matrix \mathbf{F}
- 20: **end while**
- 21: **end while**

IV. EXPERIMENTAL RESULTS

A. Experiment Setting

Datasets: We evaluate and compare our proposed framework with the state-of-the-art baselines on four datasets. The first is Animal with Attributes(AwA) [17], the second is Caltech-UCSD Birds-200-2001 (CUB-200) [36], the third is SUN scene recognition dataset(SUN) [28], and the last is aPascal-aYahoo(aPY) [6]. The AwA dataset contains 50 classes, along with 85 attributes. Similar as many works, we use 40 of the classes in the training data and the remaining 10 classes in the test data. The CUB dataset contains 200 categories of bird species with a total of 11,788 images, and each category is described by an attribute vector of 312 dimensions. We follow [2] and use the 150/50 split as the training/test classes. The SUN dataset has 717 scenes from abbey to zoo, and the attributes have 102 dimensions. Similar as [16], we split this dataset into 707 classes in the training data and the remaining 10 classes as the test data. The aPY dataset contains 20 kinds of objects from VOC challenge and related 12 classes from Yahoo image search engine, with attribute vectors of 64 dimensions. Following the standard setting, we use the aPascal subset as the source classes and the aYahoo subset as the target classes. After splitting the train/test set as aforementioned, we also divide the training set into training and validation set for hyper-parameter selection.

Image/Class input representation: For each image from all these datasets, we use the pretrained VGG-19 network as feature extractor, and use the output of its top fully-connected layer as raw input representation, which has 4096 dimensions [33]. For the representation of each class, we utilize the default class attribute features provided by the original datasets directly, which follows the standard settings for zero-shot problem [37].

Implementation details: In the VAE part, we build the encoder using a Gaussian multi-layer perceptron with three layers, and the dimension of the output layer is set as 200. We use tanh as the nonlinear activation function and the dropout rate is set as 0.8 to avoid overfitting. The decoder has a similar structure as the encoder part. The s-DBM part has four layers, and the dimension of the intermediate layers is set constantly as 200 for CUB dataset and 100 for the remaining three datasets. The effectiveness is measured using Top-1 classification accuracy [37].

B. Baselines

The performance of our proposed EBZL method is compared with a variety of state-of-the-art baselines, as shown in Table. 1. The hyphens in Table. 1 indicate that the compared methods were not tested on the corresponding datasets in the original papers and we could not find their publicly available implementations. For all these methods, we use the same set of input image features: the fc7 features of VGG-19 net as stated before. And the classes are all represented using the same attribute vectors.

To help us analyze the experiment results, a detailed description of each baseline has been given in the related work

Types	Method	AwA	CUB-200	SUN	aPY
Embedding-based	ESZSL [31]	49.3	27.27	61.53	/
Embedding-based	SP-AEN [4]	58.5	55.4	59.2	24.1
Basis-based	DAP [17]	46.1	40	39.39	33.8
Basis-based	deep RIS [23]	62.7	32.3	64.15	26.7
Basis-based	deep RULE [23]	77.6	56.3	78.83	44.17
Basis-based	SMS [13]	78.47	/	82.00	39.03
Basis-based	SCZL [3]	72.9	54.7	62.7	/
Sample Synthesize-based	SSZL [11]	82.67	55.75	85	54.04
Sample Synthesize-based	FZLC [21]	82.12	44.9	80.5	42.25
	EBZL	84.2	58.7	87.72	56.17

TABLE I: Summary of previous methods. Typical models are grouped in category and are compared in term of whether they can solve those three problems addressed in the main text.

part. Our proposed method is different from them by using a s-DBM as the energy function to model the joint probability, and is an extension of traditional embedding-based method.

C. Overall Comparison

The overall experimental results of our proposed model and various baselines are shown in Table. 1. As we test the results with public datasets, the results of most baselines are got directly from the original paper. The results clearly shows that our proposed method consistently performs the best on the four datasets. Specifically, our method EBZL performs better than the best baseline by 1.53%, 2.4%, 5.72%, 2.13% on each dataset respectively.

We experiment extensively to investigate the influence of many factors in our proposed method and baselines. The results suggest that the following three factors have a remarkable influence on the effectiveness of various methods for zero-shot problems, and show the advantage of our method over previous ones.

The semantic integrity of extracted features. Our model uses a variational autoencoder to ensure that the hidden space F preserve enough low-level information, and adopt bidirectional connections in the s-DBM part to prevent semantic loss. Samples-Synthesizing-based methods circumvent this problem by training the model in the supervised manner. However, methods like RIE and RULE did not take recovery loss into consideration, they just train the model in the classical bottom-up manner and apply it directly to unseen classes. As shown in Table. 1, our model, along with SSZL and FZLC, outperform most of baselines with a large margin on all four datasets.

The methods in modeling the relationship. Our model proposes to use energy model to represent the joint distributions, making the model much more expressive than traditional models, as energy function can model all sorts of probability distributions. FZLC represents each class as a mere point in the mediate space. SP-AEN and SCZL both model the relationship in a bottom-up manner, from the image to one certain class. They build a class-specific classifier for each class, which is hard to be generalized to other classes. As shown in Table. 1, our model, which tries to model the relationship between the

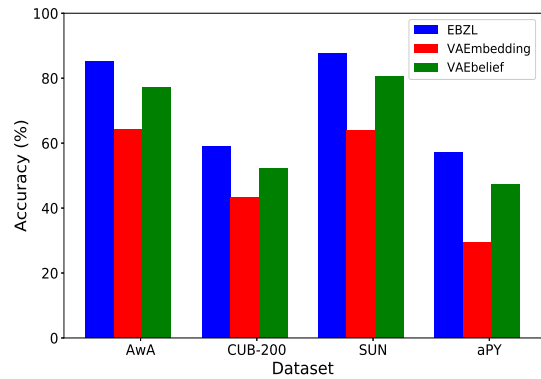


Fig. 3: Comparison of our EBZL with VAEmbedding and VAEbelief. From left to right, are the result got on datasets AWA, CUB-200, SUN and aPY separately.

image space and class space using distributions, gets the best performance.

The methods in utilizing attributes Our method takes each attribute vector as a whole to model its relationship with images, while methods such as SCZL, DAP and deep RIS assume that each dimension of attribute vectors is independent. SCZL assumes neighboring information in the attribute space can be applied to the model parameter space and synthesize classifier based on that. Deep RIE makes classification on each attribute separately, and gets the classifier by finding the class whose attribute vector is most close to the synthesized attribute vector. These strong assumptions result in their inferior performance compared to ours, as shown in Table. 1.

D. Detailed Analysis

Besides overall comparison between our proposed model with the state-of-the-art methods, we give a detailed analysis of our proposed method, including the effectiveness of model components, the parameter sensitivity analysis and model result visualization.





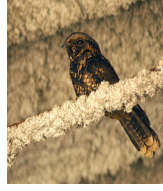



CUB	Ground truth	Class	Prob	SUN	Ground truth	Class	Prob
	Laysan Albatross	Laysan Al	0.3216		Car interior	Car interior	0.5531
		Sooty Al	0.2531		Backseat	Backseat	0.2281
		Crest Au	0.1613		Living Room	Living Room	0.0956
		Black footed Al	0.1459		Computer Room	Computer Room	0.0643
		Least Au	0.1181		Shoe Shop	Shoe Shop	0.0589
					Mineshaft	Mineshaft	
	Eastern Towhee	Eastern_To	0.2723		Cafeteria	Cafeteria	0.4136
		Black billed Cu	0.2038		Bar	Bar	0.3605
		Gray Ca	0.1954		Lab Classroom	Lab Classroom	0.1408
		Northern Fl	0.1697		Game Room	Game Room	0.0562
		Indigo Bu	0.1588		Assembly Line	Assembly Line	0.0289
	Chuck will Widow	Chuck will Wi	0.3905		Alley	Alley	0.4814
		Yellow bellied Fl	0.1773		Bazaar	Bazaar	0.2462
		Boblink	0.1593		Cloister	Cloister	0.1124
		Rusty Bl	0.1487		Castle	Castle	0.0937
		Eastern To	0.1242		Barndoor	Barndoor	0.0663
	Red faced Cormorant	red faced Co	0.2821		Lake Natural	Lake Natural	0.5249
		Pelagic Co	0.2142		Bayou	Bayou	0.2397
		Groove billed An	0.1875		Athleti Field	Athleti Field	0.1065
		Eared Gr	0.1766		Creek	Creek	0.0754
		Brandit Co	0.1396		Outhouse	Outhouse	0.0535

TABLE II: Images and their predicted classes with top-5 predicted joint probabilities(normalized).

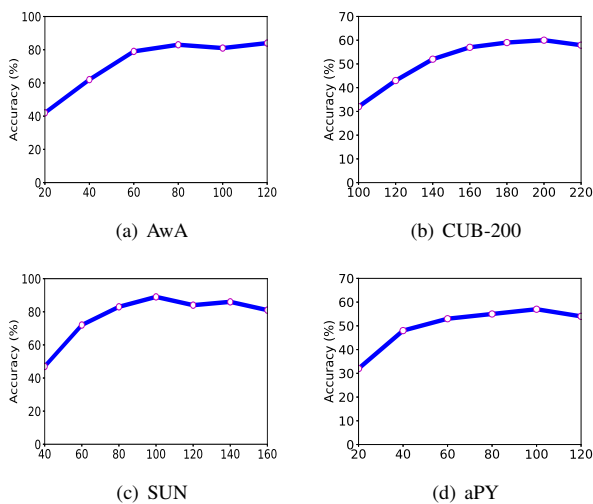


Fig. 4: Sensitivity curve of our model on the number of hidden units in each layer. X-axis denotes the number of hidden units.

Evaluation on model components. To show the effectiveness of the proposed s-DBM part of our model, we design two baseline models: VAEmbedding and VAEbelief:

- VAEmbedding: it combines a VAE part and two mapping functions, which maps the attribute vector and the features extracted by VAE to the latent embedding space. The correspondence between an image and a class is measured by their distance in the latent space.
- VAEbelief: it is composed of a VAE and a deep belief network. Its structure is similar to our model, and deep belief net is also used to measure the joint probability of an input image and an attribute vector.

As shown in Fig.3, our proposed EBZL model and VAEbelief outperform VAEmbedding with a large margin, validating the soundness of using energy function to model the correspondence between an image and a class. The difference between VAEbelief and our proposed method lies in that deep belief net has directed connections in its intermediate layers, while our s-DBM part has undirected connections to help to reduce semantic loss. As shown in Fig. 2, our model outperforms

VAEbelief with a reasonable margin, with an improvement of near 10% on each dataset. This experimental finding further validates the power of s-DBM with bidirectional connections as an energy function to model joint distributions.

Parameter sensitivity. Here, we evaluate the performance of EBZL with respect to different numbers of hidden units in each energy-function layer. The experimental results are shown in Fig. 4 on different datasets. For the dataset AWA, as we increase the dimension of hidden layers from 20 to 80, the performance increases quickly as we have more capacity in the modeling process. However, when the hidden dimension exceeds 100, the performance drops. For the remaining datasets, the sensitivity curves show similar trends, but the detailed hidden dimension varies.

Results demonstration. In Table. II, we provide some qualitative results of our method from dataset CUB and SUN. The left part of each row shows an test image from CUB with its corresponding class is shown in the ground truth column. Five classes with the top-5 predicted energy values are given in the Class column and their corresponding energy values are normalized and shown in the Prob column. The experimental results clearly show that our proposed method could accurately classify the test data.

V. CONCLUSION

In this paper, we proposed a new energy-based method named EBZL for zero-shot learning problem, which used joint probability to align the input image and class attribute spaces with bidirectional connections. It is much more expressive than previous embedding-based methods, which embed each class as a point and perform classification according to distances in that embedding space. Specifically, we trained a VAE as a feature extractor to build a discriminative feature space and modified traditional DBM into s-DBM as the energy-function part. Due to the expressive power of energy model and the bidirectional connections in s-DBM, our method circumvented semantic loss problem and outperformed all baselines with a reasonable margin. In the future, we would like to explore the possibility of using deep reinforcement learning techniques for better modeling the energy model to enhance the ZSL performance.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China(Grant No. U1605251 and Grant No. 61602147)

REFERENCES

- [1] Zeynep Akata, Mateusz Malinowski, Mario Fritz, and Bernt Schiele. Multi-cue zero-shot learning with strong supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, year=2016, pages=59-68, organization=IEEE.
- [2] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages=819-826, year=2013, organization=IEEE.
- [3] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, year=2016, pages=5327-5336, organization=IEEE.
- [4] Long Chen, Hanwang Zhang, Jun Xiao, Wei Liu, and Shih-Fu Chang. Zero-shot visual recognition using semantics-preserving adversarial embedding network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages=1043-1052, year=2018, organization=IEEE.
- [5] Mohamed Elhoseiny, Babak Saleh, and Ahmed Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *Proceedings of the IEEE Conference on Computer Vision (ICCV)*, pages=2584-2591, year=2013, organization=IEEE.
- [6] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages=1778-1785, year=2009, organization=IEEE.
- [7] Yanwei Fu, Timothy M Hospedales, Tao Xiang, and Shaogang Gong. Learning multimodal latent attributes. *IEEE transactions on pattern analysis and machine intelligence*, 36(2):303-316, 2014.
- [8] Zhenyong Fu, Tao Xiang, Elyor Kodirov, and Shaogang Gong. Zero-shot object recognition by semantic manifold distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2635-2644. IEEE, 2015.
- [9] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. In *Proceedings of the 9th International Conference on Artificial Neural Networks: ICANN '99*, pages 850-855, 1999.
- [10] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 580-587. IEEE, 2014.
- [11] Yuchen Guo, Guiguang Ding, Jungong Han, and Yue Gao. Synthesizing samples fro zero-shot learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, pages 1774-1780. IJCAI, 2017.
- [12] Yuchen Guo, Guiguang Ding, Jungong Han, and Yue Gao. Zero-shot learning with transferred samples. *IEEE Transactions on Image Processing*, 26(7):3277-3290, 2017.
- [13] Yuchen Guo, Guiguang Ding, Xiaoming Jin, and Jianmin Wang. Transductive zero-shot recognition via shared model space learning. In *Proceedings of The Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, volume 3, page 8, 2016.
- [14] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tieyan Liu, and Wei-Ying Ma. Dual learning for machine translation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 820-828, 2016.
- [15] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527-1554, 2006.
- [16] Dinesh Jayaraman and Kristen Grauman. Zero-shot recognition with unreliable attributes. In *Advances in neural information processing systems (NIPS)*, pages 3464-3472, 2014.
- [17] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453-465, 2014.
- [18] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (AAAI-2008)*, volume 1, page 3, 2008.
- [19] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and

- F Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.
- [20] Ming Liang and Xiaolin Hu. Recurrent convolutional neural network for object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3367–3375. IEEE, 2015.
- [21] Yang Long, Li Liu, Ling Shao, Fumin Shen, Guiguang Ding, and Jungong Han. From zero-shot learning to conventional supervised classification: Unseen visual data synthesis. pages 1627–1636, 2017.
- [22] Hongyuan Mei and Jason M Eisner. The neural hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems (NIPS)*, pages 6757–6767, 2017.
- [23] Pedro Morgado and Nuno Vasconcelos. Semantically consistent regularization for zero-shot recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6060–6069. IEEE, 2017.
- [24] Tu Dinh Nguyen, Dinh Phung, Viet Huynh, and Trung Le. Supervised restricted boltzmann machines. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence (UAI)*.
- [25] Tu Dinh Nguyen, Truyen Tran, Dinh Phung, and Svetha Venkatesh. Latent patient profile modelling and applications with mixed-variate restricted boltzmann machine. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 123–135. Springer, 2013.
- [26] Tu Dinh Nguyen, Truyen Tran, Dinh Phung, and Svetha Venkatesh. Learning sparse latent representation and distance metric for image retrieval. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2013.
- [27] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.
- [28] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2751–2758. IEEE, 2012.
- [29] Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. Variational autoencoder for deep learning of images, labels and captions. In *Advances in neural information processing systems (NIPS)*, pages 2352–2360, 2016.
- [30] Ruizhi Qiao, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel. 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 (CVPR)*, pages 2249–2257. IEEE, 2016.
- [31] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning (ICML)*, pages 2152–2161, 2015.
- [32] Ruslan Salakhutdinov and Hugo Larochelle. Efficient learning of deep boltzmann machines. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTAS)*, pages 693–700, 2010.
- [33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.
- [34] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. Lstm neural networks for language modeling. In *Thirteenth Annual Conference of the International Speech Communication Association*, pages 194–197, 2012.
- [35] Lorenzo Torresani, Martin Szummer, and Andrew Fitzgibbon. Efficient object category recognition using classemes. In *European conference on computer vision (ECCV)*, pages 776–789. Springer, 2010.
- [36] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.
- [37] Wenlin Wang, Yunchen Pu, Vinay Kumar Verma, Kai Fan, Yizhe Zhang, Changyou Chen, Piyush Rai, and Lawrence Carin. Zero-shot learning via class-conditioned deep generative models. In *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 4211–4218, 2018.
- [38] Laurent Younes. Parametric inference for imperfectly observed gibbsian fields. *Probability theory and related fields*, 82(4):625–645, 1989.
- [39] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2021–2030. IEEE, 2017.
- [40] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via joint latent similarity embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6034–6042. IEEE, 2016.
- [41] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, and Ahmed M. Elgammal. Imagine it for me: Generative adversarial approach for zero-shot learning from noisy texts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1004–1013. IEEE, 2018.