Information Bottleneck Constrained Latent Bidirectional Embedding for Zero-Shot Learning

Yang Liu, Lei Zhou, Xiao Bai, Lin Gu, Tatsuya Harada, Jun Zhou

1 Beihang University, Beijing, China
2 RIKEN AIP, Tokyo, Japan
3 Griffith University, Nathan, Australia

Abstract

Zero-shot learning (ZSL) aims to recognize novel classes by transferring semantic knowledge from seen classes to unseen classes. Though many ZSL methods rely on a direct mapping between the visual and the semantic space, the calibration deviation and hubness problem limit the generalization capability to unseen classes. Recently emerged generative ZSL methods generate unseen image features to transform ZSL into a supervised classification problem. However, most generative models still suffer from the seen-unseen bias problem as only seen data is used for training. To address these issues, we propose a novel bidirectional embedding based generative model with a tight visual-semantic coupling constraint. We learn a unified latent space that calibrates the embedded parametric distributions of both visual and semantic spaces. Since the embedding from high-dimensional visual features comprise much non-semantic information, the alignment of visual and semantic in latent space would inevitably be deviated. Therefore, we introduce information bottleneck (IB) constraint to ZSL for the first time to preserve essential attribute information during the mapping. Specifically, we utilize the uncertainty estimation and the wake-sleep procedure to alleviate the noises and improve model abstraction capability. We evaluate the learned latent features on four benchmark datasets. Extensive experimental results show that our method outperforms the state-of-the-art methods in different ZSL settings on most benchmark datasets. The code will be available at https://github.com/osierboy/IBZSL.

Introduction

Thanks to the abundant human annotated data, deep learning has achieved great success. However, labeling large-scale data is time consuming and expensive. Inspired by the human’s remarkable ability in recognizing instances of unseen classes solely based on class descriptions, zero-shot learning (ZSL) was proposed as an image classification setting to mimic the human learning process (Lampert, Nickisch, and Harmeling 2009). Given the semantic descriptions of both seen and unseen classes but only the seen class training images, ZSL aims to classify test images of unseen classes.

Most early ZSL methods learn a direct or indirect mapping between the visual space and the semantic space (Akata et al. 2013; Frome et al. 2013; Romera-Paredes and Torr 2015; Akata et al. 2015b; Xian et al. 2016; Guo et al. 2016; Kodirov, Xiang, and Gong 2017). However, the performance of these methods is often poor on the more challenging generalized ZSL (GZSL) setting where the test images belong to both seen and unseen classes. The reason is that the embedding model is learned only from seen classes, which leads to a bias towards predicting seen classes. To address this issue, more recent approaches (Xian et al. 2018b; Mishra et al. 2018; Wang et al. 2018; Li et al. 2019; Schonfeld et al. 2019; Ma and Hu 2020; Verma, Brahma, and Rai 2020; Yu et al. 2020) utilize generative models, e.g., generative adversarial networks (GAN) (Goodfellow et al. 2014) or variational autoencoders (VAE) (Kingma and Welling 2013), to generate synthetic features of unseen classes. This transfers
the ZSL task to a supervised classification problem. Since GAN-based loss functions are unstable in training, VAE-based methods (Schonfeld et al. 2019; Ma and Hu 2020) were developed to tackle this problem and achieved better performance. However, most of these generative models still suffer from the deviation between generated features and unseen classes due to the lack of tight visual-semantic coupling.

Since high-dimensional visual features contain non-semantic information which is redundant for semantic discrimination (Tong et al. 2019; Han, Fu, and Yang 2020; Shen, Qin, and Huang 2020), it is difficult to well align the semantic categories to the centers of visual sample distributions when mapping the semantic features to the visual space. This causes a calibration deviation problem as illustrated in Figure 1. In addition, when high-dimensional visual features are mapped to a low-dimensional semantic space, the shrink of feature space would aggravate the hubness problem that some instances in the high-dimensional space becomes the nearest neighbors of a large number of instances (Radovanovic, Nanopoulos, and Ivanovic 2010). To address the above challenges, we propose an information bottleneck (IB) (Tishby, Pereira, and Bialek 2000) constrained bidirectional embedding based generative model which utilizes advantages of both embedding model and generative model to align visual and semantic distributions in a unified latent space. As shown in Figure 2, our proposed method first learns a latent bidirectional embedding via a modified VAE network. To facilitate the distribution alignment, the redundant non-semantic information in the visual space should be discarded to preserve the attributed related part when it is flowing to the latent space. To achieve this, we design an IB loss on the latent bidirectional embedding to impose the mutual information relationships between feature spaces. Due to the widely existence of noises such as the labeling noise (Kunran Xu and Gu 2020), the human annotated semantics are insufficient to fully describe the diversified visual samples (Ding and Liu 2019). The deviation between visual and semantic distributions will accumulate during the embedding process. Therefore, we learn the bias of the original visual distribution by introducing an uncertainty estimation technique (Kendall and Gal 2017) to alleviate the influence of noises. Since one semantic class may correspond to a variety of visual samples, we also propose a bias passing mechanism to share this variety bias to the latent semantic distribution to benefit the distribution alignment. Since VAE does not incorporate the generated samples for learning, the latent features generated by VAE are largely randomized and uncontrollable (Hu et al. 2017, 2018). Therefore, we introduce a wake-sleep procedure (Hinton et al. 1995) that uses both real and generated data for joint training to improve the model representation and abstraction capability. Finally, with the generated latent features, we are able to solve the ZSL as a supervised clas-
sification problem.

The contributions of this paper are as follows. (1) We propose a novel ZSL method based on information bottleneck (IB) constrained generative model with a tight visual-semantic bidirectional embedding. The IB loss minimizes the non-semantic information when embedding the visual domain to latent space. To the best of our knowledge, this is the first work that adopts the IB theory in ZSL. (2) We exploit the data uncertainty estimation technique for the first time in ZSL to learn the bias of visual distribution and design a bias passing mechanism, which alleviates the noises and gap between visual features and human annotated semantics. (3) We train the model on both real and generated data with a wake-sleep training mechanism to improve the model representation and abstraction capability via a VAE model. (4) Extensive experiments on both GZSL and conventional ZSL show the superiority of our method.

Related Works

In this section, we review related works on ZSL and specifically focus on GZSL.

Embedding Models

Embedding models for ZSL focus on learning a direct or indirect mapping between visual and semantic spaces to transfer semantic knowledge from seen classes to unseen classes. There are three typical embedding strategies. The earliest methods learned the mapping function from the visual space to the semantic space, which include, for example DAP and IAP (Lampert, Nickisch, and Harmeling 2013), ALE (Akata et al. 2015a), DeViSE (Frome et al. 2013) and ESZSL (Romera-Paredes and Torr 2015). To alleviate the severe hubness problem caused by embedding from the high-dimensional visual space to the low-dimensional semantic space, reverse mapping from the semantic space to the visual space was proposed for the nearest neighbor classification in the visual space (Changpinyo, Chao, and Sha 2017; Zhang, Xiang, and Gong 2017). Some models such as SSE (Zhang and Saligrama 2015), SYNC (Changpinyo et al. 2016) and BiDiLEL (Wang and Chen 2017) explore the idea of embedding both visual and semantic features into a common intermediate space. Though these methods perform well in the ZSL setting, their performance deteriorates on GZSL setting since there are only seen class features for model training.

Generative models

Recently, abundant generative models (Guo et al. 2017; Chen et al. 2018; Felix et al. et al. 2018; Kumar Verma et al. 2018; Xian et al. 2018b; Zhu et al. 2018; Li et al. 2019; Schonfeld et al. 2019; Ma and Hu 2020; Keshari, Singh, and Vatsa 2020) have been used for ZSL. f-CLSWGAN (Xian et al. 2018b) uses a Wasserstein GAN (Arjovsky, Chintala, and Bottou 2017; Gulrajani et al. 2017) to synthesize samples. LisGAN (Li et al. 2019) exploits the multi-view of each class to guide the sample generation. Due to the hardness of training GAN based models, CADA-VAE (Schonfeld et al. 2019) adopts a cross-aligned VAE to align the visual and semantic distributions in a latent space. More recently, a new flow-based generative model (Shen, Qin, and Huang 2020) was introduced to ZSL which utilizes an invertible generative flow networks to generate distinguishable samples.

Although generative models have been quite successful in GZSL, feature generation for unseen classes still needs tight visual-semantic coupling constraints to alleviate the deviation. Our proposed method combines the advantages of both embedding and generative model for an accurate alignment of visual-semantic distributions while generating discriminative image features simultaneously.

Proposed Method

In this section, we first define the problem setting, notations and then present the details of each module of our method.

Problem Setting and Notations

The GZSL problem is defined as follows. Let $\mathcal{X}^S$ and $\mathcal{X}^U$ denote the image feature sets of seen classes and unseen classes respectively, $\mathcal{X} = \mathcal{X}^S \cup \mathcal{X}^U$, $\mathcal{S} = \{(x, y, c(y)) | x \in \mathcal{X}^S, y \in \mathcal{Y}^S, c(y) \in \mathcal{C}^S\}$ denotes the training set, where $x \in \mathbb{R}^D$ are image features extracted by a plain CNN model, $y$ are the seen class labels which are available during training and $c(y) \in \mathbb{R}^K$ are attribute features. The auxiliary training set is $\mathcal{U} = \{(u, c(u)) | u \in \mathcal{Y}^U, c(u) \in \mathcal{C}^U\}$, where $u$ denote unseen class labels. The seen classes and unseen classes are disjoint, i.e., $\mathcal{Y}^S \cap \mathcal{Y}^U = \emptyset$. Here, $\mathcal{C} = \mathcal{C}^S \cup \mathcal{C}^U$ is used to transfer information between seen classes and unseen classes. $\mathcal{C}$ can be human-annotated attributes (Xian et al. 2018b) or articles describing the classes (Zhu et al. 2018). In the conventional ZSL, the task is to learn a classifier $f^{\text{ZSL}} : \mathcal{X}^U \rightarrow \mathcal{Y}^U$. However, in more realistic and challenging setup of GZSL, the aim is to learn a classifier $f^{\text{GZSL}} : \mathcal{X} \rightarrow \mathcal{Y}^U \cup \mathcal{Y}^S$.

The architecture of the proposed model is shown in Figure 2. It consists of two sets of latent embedding VAEs with $(E_{v \rightarrow l}, D_{l \rightarrow a})$ and $(E_{a \rightarrow l}, D_{l \rightarrow v})$. These two sets of VAEs share the same latent space $l$, $E_{a \rightarrow l}$ maps visual space $v$ to latent space $l$, and $D_{l \rightarrow a}$ maps latent space $l$ to semantic space $a$. $E_{v \rightarrow l}$ maps semantic space $a$ to latent space $l$, and $D_{l \rightarrow v}$ maps latent space $l$ to visual space $v$.

Latent Bidirectional Embedding

The goal of our model is to learn a latent space that can accurately align visual and semantic distributions. We first learn a Visual to Semantic (VS) network $\mathcal{V} = E_{a \rightarrow l} \circ D_{l \rightarrow a} : \mathbb{R}^D \rightarrow \mathbb{R}^K$ to project the visual features through latent space into semantic space. The latent embedding model is shown in Figure 3. Because there may be inherent noise in the visual features (Chang et al. 2020). To reduce the impact of

\begin{align*}
E_{v \rightarrow l} : \mathbb{R}^D & \rightarrow \mathbb{R}^K \\
D_{l \rightarrow v} : \mathbb{R}^K & \rightarrow \mathbb{R}^D \\
E_{a \rightarrow l} : \mathbb{R}^K & \rightarrow \mathbb{R}^K \\
D_{l \rightarrow a} : \mathbb{R}^K & \rightarrow \mathbb{R}^K
\end{align*}
data uncertainty, we define the latent representation \( z_i^{(v)} \) embedded from each visual sample \( \mathbf{x}_i \) as a Gaussian distribution:

\[
p(z_i^{(v)} | \mathbf{x}_i) = \mathcal{N}(z_i^{(v)}; \mu_i^{(v)}, \sigma_i^2 \mathbf{I})
\]

where \( \mu_i^{(v)} \) and \( \sigma_i^2 \) are the mean and variance of the Gaussian distribution learned by the encoder \( E_{v \rightarrow l} : \mu_i^{(v)} = \tilde{E}_{v \rightarrow l, \phi_1} \mathbf{x}_i, \log \sigma_i^2 = \tilde{E}_{v \rightarrow l, \phi_2} \mathbf{x}_i \), where \( \phi_1 \) and \( \phi_2 \) refer to the model parameters. The re-parameterization trick (Kingma and Welling 2013) is used to keep gradients of the model as usual. With this trick, we generate the latent sampling representation \( s_i^{(v)} \) as

\[
s_i^{(v)} = \mu_i^{(v)} + \epsilon \sigma_i, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I})
\]

where \( \epsilon \) is a random noise.

Then, \( \tilde{c}(y_i) = D_{l \rightarrow a}(s_i^{(v)}) \) projects the latent feature \( s_i^{(v)} \) into semantic space, i.e., the mapping of a visual sample \( \mathbf{x}_i \) is calculated as \( \text{VS}(\mathbf{x}_i) \):

\[
\text{VS}(\mathbf{x}_i) = \tilde{c}(y_i) = D_{l \rightarrow a}(s_i^{(v)}) = E_{v \rightarrow l} \circ D_{l \rightarrow a}(\mathbf{x}_i)
\]

The affinity between \( \text{VS}(\mathbf{x}_i) \) and the \( y_i \)-th attribute feature \( c(y_i) \) could be measured by their inner product \( \text{VS}(\mathbf{x}_i)^T \mathbf{c}(y_i) \). Then the probability of \( \mathbf{x}_i \) belong to the \( y_i \)-th category in semantic space can be calculated as:

\[
p^A(y_i | \mathbf{x}_i) = \frac{\exp(\text{VS}(\mathbf{x}_i)^T \mathbf{c}(y_i))}{\sum_{y \in Y} \exp(\text{VS}(\mathbf{x}_i)^T \mathbf{c}(y))}
\]

Then the Semantic Cross-Entropy (SCE) loss can be written as:

\[
L_{SCE} = - \sum_i \log p^A(y_i | \mathbf{x}_i)
\]

Similarity, we learn a Semantic to Visual (SV) network \( SV = E_{a \rightarrow l} \circ D_{l \rightarrow v} : \mathbb{R}^K \rightarrow \mathbb{R}^D \), which first projects the semantic feature \( c(y_i) \) to latent space as \( \mu_i^{(a)} \), then projects \( \mu_i^{(a)} \) to visual space as the generated visual prototype \( \tilde{x}(y_i) \) for the \( y_i \)-th category:

\[
SV(c(y_i)) = \tilde{x}(y_i) = D_{l \rightarrow v}(\mu_i^{(a)}) = E_{a \rightarrow l} \circ D_{l \rightarrow v}(c(y_i))
\]

The probability of \( \mathbf{x}_i \) belong to the \( y_i \)-th category in visual space is calculated as:

\[
p^V(\mathbf{x}_i | y_i) = \frac{\exp(\mathbf{x}_i^T \mathbf{c}(y_i))}{\sum_{y \in Y} \exp(\mathbf{x}_i^T \mathbf{c}(y))}
\]

Then the Visual Cross-Entropy (VCE) loss is:

\[
L_{VCE} = - \sum_i \log p^V(y_i | \mathbf{x}_i)
\]

The total Cross-Entropy (CE) loss is as follow:

\[
L_{CE} = L_{SCE} + L_{VCE}
\]

In order to learn an accurate latent bidirectional embedding, we perform center calibration for each category. Such a structured objective requires the center embedding of \( \mathbf{x}_i \) being closer to the latent embedding of its groundtruth \( c(y_i) \) than other classes, the Center Calibration (CC) is defined as:

\[
L_{CC} = \sum_{i, y} \left[ \Delta + d(E_{v \rightarrow l, \phi_1}(\mathbf{x}_i), E_{a \rightarrow l}(c(y_i))) \right] + d(E_{v \rightarrow l, \phi_1}(\mathbf{x}_i), E_{a \rightarrow l}(c(y_i)))
\]

where \( d(\cdot, \cdot) \) denotes a certain distance metric. Here, we utilize the Euclidean distance in the experiments. \( \Delta > 0 \) is a margin to make \( L_{CC} \) more robust.

**Feature Generation**

For each category \( y_i \), there could be many visual samples \( \mathbf{x} \), but the semantic description \( c \) of each category is unique. Thus, this unique semantic attribute \( c \) is insufficient to fully describe the variety of visual samples. Therefore, we assume the latent semantic distribution similar to the Gaussian distribution of latent visual features in equation (1). To adapt to this task, we use two sets of encoder-decoder structures. \( E_{v \rightarrow l} \) encodes the visual features \( \mathbf{x}_i \) to a Gaussian distribution \( \mathcal{N}(\mu_i^{(v)}, \sigma_i^2) \) in the latent space, and \( E_{a \rightarrow l} \) encodes the semantic feature \( c(y_i) \) to the center \( \mu_i^{(a)} \) of category \( y_i \).

Since the latent semantic Gaussian distribution \( \mathcal{N}(\mu_i^{(a)}, \sigma_i^2) \) should be consistent with the latent visual distribution, we design a bias passing mechanism to share the bias for the latent semantic distribution. Then we use the decoders \( D_{l \rightarrow a} \) to decode \( \mu_i^{(v)} \) or \( \mu_i^{(a)} \) to semantic feature \( \tilde{c}(y_i) \), and use \( D_{l \rightarrow v} \) to decode \( s_i^{(v)} \sim \mathcal{N}(\mu_i^{(v)}, \sigma_i^2) \) or \( s_i^{(a)} \sim \mathcal{N}(\mu_i^{(a)}, \sigma_i^2) \) to visual feature \( \tilde{x}_i \). Finally, the loss for the modified VAE can be written as:

\[
L_{VAE} = E_{q_\phi(s^{(v)} | \mathbf{x})} [\log p^A(c | s^{(v)})]
+ E_{q_\phi(s^{(a)} | c)} [\log p \theta_1(c | s^{(a)})]
+ E_{q_\phi(s^{(v)} | c)} [\log p \theta_2(c | s^{(v)})]
+ \beta D_{KL}(q_\phi(s^{(v)} | \mathbf{x}) || \mathcal{N}(\mu^{(a)}, \mathbf{I}))
\]
where, \( \phi \) refers to the parameters of \( E_{v \rightarrow l} \) and \( E_{a \rightarrow l} \), \( \theta_1 \) and \( \theta_2 \) refer to the parameters of \( D_{l \rightarrow a} \) and \( D_{l \rightarrow v} \), respectively.

**Information Bottleneck Constraint**

In our method, information is gradually disentangled from the visual space through the latent space to the semantic space. The semantic feature \( c \) is related disentangled attribute information while the visual feature \( x \) has high-dimensional entangled non-semantic information. Therefore, we hope that the latent feature \( s \) should contain as much semantic information of \( c \) as possible while discard the redundant non-semantic information of \( x \). In information theory, the dependence between two random variables could be measured by mutual information \( I(\cdot;\cdot) \). As illustrated in Figure 1, we maximize the mutual information between the semantic space and the latent space \( I(s;c) \) and minimize the mutual information between the visual space and the latent space \( I(s;x) \). We define a information bottleneck (IB) (Tishby, Pereira, and Bialek 2000) to constrain the information relationship between spaces:

\[
\max I(s;c) - \eta I(s;x)
\]

Since \( s \) may be sampled from different distributions like \( N(\mu^v, \sigma^2) \) or \( N(\mu^a, \sigma^2) \) in our model, we resample \( s^* \sim N(\mu^*, \sigma^2) \), where \( \mu^* = \alpha \mu^v + (1 - \alpha) \mu^a \) with a uniform distribution \( \alpha \sim U(0,1) \).

Since the VAE model does not utilize the generated samples for training, the latent features generated are largely randomized and uncontrollable. Inspired by the work of Hu et al. (Hu et al. 2018), we train the modified VAE model in a wake-sleep procedure, using real data and generated data for joint training. The extended wake-sleep procedure is shown in Figure 4. In the wake phase, we use real visual data \( x \) to train the feature representation capability of the model. In the sleep phase, we use generated data \( \bar{x} \) to train the abstraction capability of the model. So that the model can generate disentangled latent features. The wake-sleep information bottleneck constraint is as follow:

\[
\max [I(s^*;c) - \eta I(s^*;x)] + \lambda [I(\bar{s}^*;c) - \eta I(\bar{s}^*;\bar{x})]
\]

where \( \bar{s}^* \) is the latent embedding representation of \( \bar{x} \).

Since the information bottleneck with high dimension is intractable to calculate, we follow the strategy proposed by Alemi et al. (Alemi et al. 2016). The Information Bottleneck (IB) loss is shown as follow:

\[
\mathcal{L}_{IB} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[-\log q(c(y_i)D_{l \rightarrow a}(s^*_i))] + \eta D_{KL}(q_\phi(s^* | x_i, c(y_i))^{|r(z)|}
\]

where \( r(z) \) is a standard normal distribution in the experiments, \( \eta \) is initialized to \( 10^{-5} \) and changed with IB loss.

Finally, the overall loss of the proposed model is defined as:

\[
\mathcal{L} = \mathcal{L}_{VAE} + \gamma \mathcal{L}_{CE} + \delta \mathcal{L}_{CC} + \tau \mathcal{L}_{IB}
\]

where \( \gamma, \delta \) and \( \tau \) are the weighting factors of the cross entropy loss, center calibration and information bottleneck loss, respectively.

**Implementation Details**

In our modified VAE model, we utilized multilayer perceptrons to implement the encoders \( (E_{v \rightarrow l} \) and \( E_{a \rightarrow l} \)) and decoders \( (D_{l \rightarrow a} \) and \( D_{l \rightarrow v} \)). The encoders \( E_{v \rightarrow l} \) and \( E_{a \rightarrow l} \) had 1560 and 1450 hidden units, respectively. The hidden units of \( D_{l \rightarrow a} \) and \( D_{l \rightarrow v} \) were 665 and 1660, respectively. The latent embedding dimensions were 64 for AWA2 and aPY and 256 for CUB and SUN. \( \beta, \lambda, \gamma, \delta \) and \( \tau \) were set to 0.5, 0.1, 1.0, 0.1 and 1.0. The margin \( \Delta \) was set to \( 10^{-3} \). Adam optimizer (Kingma and Ba 2014) was used for training, the epoch size was 120 and the batch size was 64. After the model training, the encoders \( E_{v \rightarrow l} \) and \( E_{a \rightarrow l} \) transformed the visual features of seen classes and attribute features of unseen classes into the unified latent space. Finally, we trained a softmax classifier to classify latent features.

**Experiments**

We evaluated our framework on four widely used benchmark datasets including CUB-200-2011 (CUB) (Welinder et al. 2010), SUN attribute (SUN) (Patterson and Hays 2012), attributes Pascal and Yahoo (aPY) (Farhadi et al. 2009) and Animals with Attributes 2 (AWA2) (Xian et al. 2018a) for the GZSL. We extracted a 2,048-dimensional CNN features for images using ResNet-101 (He et al. 2016) as the visual features. The pre-defined attributes on each dataset were used as the semantic descriptors. Moreover, we adopted the Proposed Split (PS) (Xian et al. 2018a) to divide all classes into seen and unseen classes on each dataset.

The performance of our method is evaluated by per-class Top-1 accuracy. In GZSL, since the test set is composed of seen and unseen images, the Top-1 accuracy evaluated respectively on seen classes, denoted as \( S \), and unseen classes, denoted as \( U \). Their harmonic mean, defined as \( H = (2 \times S \times U) / (S + U) \) (Xian et al. 2018a), are used to evaluate the performance of GZSL.

**Comparison with the State-of-the-Art**

We selected several state-of-the-art GZSL methods for comparison, which include non-feature generation methods ALE (Akata et al. 2013), DeVise (Frome et al. 2013), ESZSL (Romera-Paredes and Torr 2015), SJE (Akata et al. 2016). The Information Bottleneck (IB) loss is shown as follows:

\[
\mathcal{L}_{IB} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[-\log q(c(y_i)D_{l \rightarrow a}(s^*_i))] + \eta D_{KL}(q_\phi(s^* | x_i, c(y_i))^{|r(z)|}
\]

where \( r(z) \) is a standard normal distribution in the experiments, \( \eta \) is initialized to \( 10^{-5} \) and changed with IB loss.

Finally, the overall loss of the proposed model is defined as:

\[
\mathcal{L} = \mathcal{L}_{VAE} + \gamma \mathcal{L}_{CE} + \delta \mathcal{L}_{CC} + \tau \mathcal{L}_{IB}
\]

where \( \gamma, \delta \) and \( \tau \) are the weighting factors of the cross entropy loss, center calibration and information bottleneck loss, respectively.
Table 1: Results of the state-of-the-arts generalized zero-shot learning.

| Methods                  | CUB   | AWA2  | SUN   | aPY   |
|--------------------------|-------|-------|-------|-------|
|                          | U   | S   | H   | U   | S   | H   | U   | S   | H   | U   | S   | H   |
| ALE (Akata et al. 2013)  | 23.7| 62.8| 34.5| 14.0| 81.8| 23.9| 21.8| 33.1| 26.3| 4.6| 73.3| 8.7 |
| DeViSE (Frome et al. 2013) | 23.8| 53.0| 32.8| 17.1| 74.7| 27.8| 16.9| 72.4| 20.9| 4.9| 76.9| 9.2 |
| ESZSL (Romera-Paredes and Torr 2015) | 12.6| 63.8| 21.0| 5.9 | 77.8| 11.0| 11.0| 72.9| 15.8| 2.4| 70.1| 4.6 |
| SJE (Akata et al. 2015b) | 23.5| 59.2| 33.6| 8.0 | 73.9| 14.4| 14.7| 30.5| 19.8| 3.7| 55.7| 6.9 |
| LATEM (Xian et al. 2016) | 15.2| 57.3| 24.0| 11.5| 77.3| 20.0| 14.7| 28.8| 19.5| 0.1| 73.0| 0.2 |
| SYNC (Changpinyo et al. 2016) | 11.5| 70.9| 19.8| 10.0| 90.5| 18.0| 7.9 | 43.3| 13.4| 7.4| 66.3| 13.3 |
| SAE (Kodirov, Xiang, and Gong 2017) | 7.8 | 54.0| 13.6| 1.1 | 82.2| 2.2 | 8.8 | 18.0| 11.8| 0.4| 80.9| 0.9 |
| SP-AEN (Chen et al. 2018) | 34.7| 70.6| 46.6| 23.3| 90.0| 37.1| 24.9| 38.6| 30.0| 13.7| 63.4| 22.6 |
| TCN (Jiang et al. 2019) | 52.6| 52.0| 52.3| 61.2| 65.8| 63.4| 31.2| 37.3| 34.0| 24.1| 64.0| 35.1 |
| TripleLoss (Cacheux, Borgne, and Crucianu 2019) | 55.8 | 52.3| 53.0| 48.5| 83.2| 61.3| 47.9| 30.4| 36.8| - | - | - |
| SE-GZSL (Kumar Verma et al. 2018) | 41.5| 53.3| 46.7| 58.3| 68.1| 62.8| 40.9| 30.5| 34.9| - | - | - |
| CV-AE-ZSL (Mishra et al. 2018) | - | - | 34.5| - | - | 51.2| - | - | 26.7| - | - | - |
| f-CLSWGAN (Xian et al. 2018b) | 43.7| 57.7| 49.7| 53.8| 68.2| 60.2| 42.6| 36.6| 39.4| - | - | - |
| LisGAN (Li et al. 2019) | 46.5| 57.9| 51.6| 54.3| 68.5| 60.6| 42.9| 37.8| 40.2| 34.3| 68.2| 45.7 |
| GDAN (Huang et al. 2019) | 39.3| 66.7| 49.5| 32.1| 67.5| 43.5| 38.1| 89.9| 53.4| 30.4| 75.0| 43.4 |
| CADA-VAE (Schonfeld et al. 2019) | 51.6| 53.5| 52.4| 55.8| 75.0| 63.9| 47.2| 35.7| 40.6| - | - | - |
| ABP (Zhu et al. 2019) | 47.0| 54.8| 50.6| 55.3| 72.6| 62.6| 45.3| 36.8| 40.6| - | - | - |
| OCD-CVAE (Keshari, Singh, and Vatsa 2020) | 44.8| 55.9| 51.3| 59.5| 73.4| 65.7| 44.8| 42.9| 43.8| - | - | - |
| Ours                     | 52.2| 56.2| 54.1| 56.0| 80.0| 65.9| 43.8| 37.8| 40.6| 34.2| 69.8| 45.9 |

Table 2: Results of conventional zero-shot learning.

| Methods                  | CUB | AWA2 | SUN | aPY |
|--------------------------|-----|------|-----|-----|
|                          | U   | S   | H   | U   |
| ALE                      | 54.9| 62.5| 58.1| 39.7|
| DeViSE                   | 52.0| 59.7| 56.5| 39.8|
| ESZSL                    | 53.9| 58.6| 54.5| 38.3|
| SJE                      | 53.9| 61.9| 53.7| 32.9|
| LATEM                    | 49.3| 55.8| 55.3| 35.2|
| SYNC                     | 55.6| 46.6| 56.3| 23.9|
| SAE                      | 33.3| 54.1| 40.3| 8.3 |
| TCN                      | 59.5| 71.2| 61.5| 38.9|
| SE-GZSL                  | 59.6| 69.2| 63.4| -   |
| CV-AE-ZSL                | 52.1| 65.8| 61.7| -   |
| f-CLSWGAN                | 57.3| -   | 60.8| -   |
| LisGAN                   | 58.8| -   | 61.7| 43.1|
| ABP                      | 58.5| 70.4| 61.5| -   |
| OCD-CVAE                 | 60.3| 71.3| 63.5| -   |
| Ours                     | 62.2| 70.1| 64.2| 43.5|

Table 2: Results of conventional zero-shot learning.

outperforms all the ten compared non-feature generation methods with a large margin for the harmonic mean results. Moreover, our method significantly improve the Top-1 accuracy on unseen classes benefit from the generated unseen class samples. Compared with the feature generation based methods, our method can also achieve the best harmonic mean results on CUB, AWA2 and aPY. Since the IB constrained bidirectional embedding between the visual space and the semantic space can preserve essential attribute information and discard the non-semantic information.

To further demonstrate the effectiveness of our method. We also compared our method under the conventional ZSL setting that the test image only belongs to unseen classes. As shown in Table 2, our proposed method achieves the best for three out of the four datasets.

Ablation Study

We conducted ablation experiments to verify the effectiveness of the proposed modules. Table 3 shows the influence of different losses. We can see that our proposed method achieves the best harmonic mean results with all the losses. Specifically, the proposed IB loss can significantly improve the performance. For the proposed wake-sleep IB constraint, we also performed ablation study with different conditions on CUB, as shown in Table 4. It can be seen that the IB loss constrained on the generated seen classes features (Sleep(S)) has significantly improved the classification accuracy of the seen classes and conventional ZSL. Accordingly, the IB loss constrained on the generated unseen classes features (Sleep(U)) also improves the result of unseen classes. Our method achieves the highest harmonic mean result under the wake-sleep IB constraint. In addition, we use $N(\mu^{(o)}, I)$ to replace the latent seman-
Table 3: Ablation study of the proposed modules.

| Loss | CUB | AWA2 | SUN | aPY |
|------|-----|------|-----|-----|
| $\mathcal{L}_{VAE}$ | $\mathcal{L}_{CE}$ | $\mathcal{L}_{CC}$ | $\mathcal{L}_{IB}$ | $\mathcal{L}_{VAE}$ | $\mathcal{L}_{CE}$ | $\mathcal{L}_{CC}$ | $\mathcal{L}_{IB}$ | $\mathcal{L}_{VAE}$ | $\mathcal{L}_{CE}$ | $\mathcal{L}_{CC}$ | $\mathcal{L}_{IB}$ | $\mathcal{L}_{VAE}$ | $\mathcal{L}_{CE}$ | $\mathcal{L}_{CC}$ | $\mathcal{L}_{IB}$ |
| U   | S   | H    | U   | S   | H    | U   | S   | H    | U   | S   | H    | U   | S   | H    | U   | S   | H    |
| 46.6 | 56.9 | 51.2 | 52.8 | 77.9 | 62.9 | 40.1 | 35.7 | 37.7 | 32.7 | 61.8 | 42.8 | 45.6 | 50.0 | 55.4 | 52.5 |
| 50.8 | 56.6 | 53.5 | 54.2 | 77.0 | 63.6 | 43.8 | 36.9 | 40.1 | 32.1 | 66.7 | 43.3 | 52.2 | 54.1 | 53.2 | 51.6 |
| 51.0 | 56.3 | 53.5 | 53.2 | 87.0 | 64.1 | 41.8 | 37.8 | 39.7 | 34.7 | 65.4 | 45.3 | 61.8 | 42.8 | 45.6 | 50.0 |
| 52.2 | 56.2 | 54.1 | 56.0 | 80.0 | 65.9 | 43.8 | 37.8 | 40.6 | 34.2 | 69.8 | 45.9 | 61.8 | 42.8 | 45.6 | 50.0 |

Table 4: Ablation study of wake-sleep IB constraint on CUB dataset. Wake phase only uses real data for training. Sleep(S/U) phase uses generated seen/unseen classes data for training.

| Wake | Sleep(S) | Sleep(U) | ZSL | U   | S   | H    |
|------|---------|---------|-----|-----|-----|------|
|      |         |         |     | 60.5 | 50.0 | 55.4 |
|      |         |         |     | 61.9 | 48.7 | 58.5 |
|      |         |         |     | 60.9 | 51.6 | 54.8 |
| ✓    | ✓       | ✓       |     | 62.2 | 52.2 | 54.1 |

Figure 5: The influence of the latent feature dimensions.

Figure 6: The influence of different numbers of generated features per seen and unseen classes.

Figure 7: The influence of different numbers of generated features per seen and unseen classes on CUB. Figure 6 reports the harmonic mean results of GZSL. We can see that the GZSL performance of our method increases with more generated unseen features in most cases and when the number of generated unseen features is twice the generated seen features, our method achieves the best result.

Visualization Result

We use the t-SNE (Maaten and Hinton 2008) to visualize our latent features used for final GZSL classification. Figure 7 shows the distributions of the latent features of 50 classes on CUB dataset. The top is the latent visual embedding features and the bottom is the latent semantic embedding features. From the almost consistent distribution, we can see our latent features can well align visual and semantic distributions.

Further Analyses

We first evaluated our method with different dimensions of latent features, as shown in Figure 5. The harmonic mean results have less fluctuation with different latent feature dimensions on four datasets. Our method achieves the best performance on CUB and SUN with the latent feature dimensions equal to 256. The best results are reached for AWA2 and aPY when the latent feature dimension is 64. We speculate the reason is that on the one hand the CUB and SUN are fine-grained datasets which need more information to distinguish. On the other hand, excessive dimensions lead to redundant information.

Then we show the influence of different numbers of generated features per seen and unseen classes on CUB. Figure 6 reports the harmonic mean results of GZSL. We can see that the GZSL performance of our method increases with more generated unseen features in most cases and when the number of generated unseen features is twice the generated seen features, our method achieves the best result.
Conclusion

In this paper, we have introduced a novel bidirectional embedding based generative model for zero-shot learning. This method learns a unified latent space to align the feature distributions of both visual domain and semantic domain. A novel information bottleneck (IB) constrained latent bidirectional embedding allows the latent features to contain more essential attributes related information while discarding non-semantic information flowed from the visual features. In addition, data uncertainty estimation and wake-sleep procedure are introduced to facilitate latent distributions alignment. The proposed method has outperformed several state-of-the-art methods in different ZSL settings in experimental comparison, showing the advantages of our approach.

References

Akata, Z.; Perronnin, F.; Harchaoui, Z.; and Schmid, C. 2013. Label-embedding for attribute-based classification. In CVPR, 819–826.

Akata, Z.; Perronnin, F.; Harchaoui, Z.; and Schmid, C. 2015a. Label-embedding for image classification. IEEE Transactions on Pattern Analysis and Machine Intelligence 38(7): 1425–1438.

Akata, Z.; Reed, S.; Walter, D.; Lee, H.; and Schiele, B. 2015b. Evaluation of output embeddings for fine-grained image classification. In CVPR, 2927–2936.

Aleimi, A. A.; Fischer, I.; Dillon, J. V.; and Murphy, K. 2016. Deep variational information bottleneck. arXiv preprint arXiv:1612.00410.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. In ICML, 214–223.

Bucher, M.; Herbin, S.; and Jurie, F. 2017. Generating visual representations for zero-shot classification. In ICCV, 2666–2673.

Cacheux, Y. L.; Borgne, H. L.; and Crucianu, M. 2019. Modeling inter and intra-class relations in the triplet loss for zero-shot learning. In ICCV, 10333–10342.

Chang, J.; Lan, Z.; Cheng, C.; and Wei, Y. 2020. Data Uncertainty Learning in Face Recognition. In CVPR, 5710–5719.

Changpinyo, S.; Chao, W.-L.; Gong, B.; and Sha, F. 2016. Synthesized classifiers for zero-shot learning. In CVPR, 5327–5336.

Changpinyo, S.; Chao, W.-L.; and Sha, F. 2017. Predicting visual exemplars of unseen classes for zero-shot learning. In ICCV, 3476–3485.

Chen, L.; Zhang, H.; Xiao, J.; Liu, W.; and Chang, S.-F. 2018. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In CVPR, 1043–1052.

Ding, Z.; and Liu, H. 2019. Marginalized latent semantic encoder for zero-shot learning. In CVPR, 6191–6199.

Farhadi, A.; Endres, I.; Hoiem, D.; and Forsyth, D. 2009. Describing objects by their attributes. In CVPR, 1778–1785.

Felix, R.; Kumar, V. B.; Reid, I.; and Carneiro, G. 2018. Multi-modal cycle-consistent generalized zero-shot learning. In ECCV, 21–37.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; and Mikolov, T. 2013. Devise: A deep visual-semantic embedding model. In NIPS, 2121–2129.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In NIPS, 2672–2680.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C. 2017. Improved training of wasserstein gans. In NIPS, 5767–5777.

Guo, Y.; Ding, G.; Han, J.; and Gao, Y. 2017. Synthesizing samples for zero-shot learning. In IJCAI, 1774–1780.

Guo, Y.; Ding, G.; Jin, X.; and Wang, J. 2016. Transductive Zero-Shot Recognition via Shared Model Space Learning. In AAAI, 3494–3500.

Han, Z.; Fu, Z.; and Yang, J. 2020. Learning the Redundancy-Free Features for Generalized Zero-Shot Object Recognition. In CVPR, 12865–12874.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In CVPR, 770–778.

Hinton, G. E.; Dayan, P.; Frey, B. J.; and Neal, R. M. 1995. The “wake-sleep” algorithm for unsupervised neural networks. Science 268(5214): 1158–1161.

Hu, Z.; Yang, Z.; Liang, X.; Salakhutdinov, R.; and Xing, E. P. 2017. Toward Controlled Generation of Text. In ICML, 1587–1596.

Hu, Z.; Yang, Z.; Salakhutdinov, R.; and Xing, E. P. 2018. On Unifying Deep Generative Models. In ICLR.

Huang, H.; Wang, C.; Yu, P. S.; and Wang, C.-D. 2019. Generative dual adversarial network for generalized zero-shot learning. In CVPR, 801–810.
Jiang, H.; Wang, R.; Shan, S.; and Chen, X. 2019. Transferable contrastive network for generalized zero-shot learning. In ICCV, 9765–9774.

Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In NIPS, 5574–5584.

Keshari, R.; Singh, R.; and Vatsa, M. 2020. Generalized Zero-Shot Learning Via Over-Complete Distribution. In CVPR, 13300–13308.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kodirov, E.; Xiang, T.; and Gong, S. 2017. Semantic autoencoder for zero-shot learning. In CVPR, 3174–3183.

Kumar Verma, V.; Arora, G.; Mishra, A.; and Rai, P. 2018. Generalized zero-shot learning via synthesized examples. In CVPR, 4281–4289.

Kunran Xu, Lai Rui, Y. L.; and Gu, L. 2020. Feature Normalized Knowledge Distillation for Image Classification. In ECCV.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2009. Learning to detect unseen object classes by between-class attribute transfer. In CVPR, 951–958.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2013. Attribute-based classification for zero-shot visual object categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence 36(3): 453–465.

Li, J.; Jing, M.; Lu, K.; Ding, Z.; Zhu, L.; and Huang, Z. 2019. Leveraging the invariant side of generative zero-shot learning. In CVPR, 7402–7411.

Ma, P.; and Hu, X. 2020. A Variational Autoencoder with Deep Embedding Model for Generalized Zero-Shot Learning. In AAAI, 11733–11740.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. Journal of Machine Learning Research 9(Nov): 2579–2605.

Mishra, A.; Krishna Reddy, S.; Mittal, A.; and Murthy, H. A. 2018. A generative model for zero shot learning using conditional variational autoencoders. In CVPRW, 2188–2196.

Patterson, G.; and Hays, J. 2012. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In CVPR, 2751–2758.

Radovanovic, M.; Nanopoulos, A.; and Ivanovic, M. 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. Journal of Machine Learning Research 11(sept): 2487–2531.

Romera-Paredes, B.; and Torr, P. 2015. An embarrassingly simple approach to zero-shot learning. In ICML, 2152–2161.

Schonfeld, E.; Ebrahimi, S.; Sinha, S.; Darrell, T.; and Akata, Z. 2019. Generalized zero-and few-shot learning via aligned variational autoencoders. In CVPR, 8247–8255.

Shen, Y.; Qin, J.; and Huang, L. 2020. Invertible zero-shot recognition flows. In ECCV.

Tishby, N.; Pereira, F. C.; and Bialek, W. 2000. The information bottleneck method. arXiv preprint physics/0004057.

Tong, B.; Wang, C.; Klinkigt, M.; Kobayashi, Y.; and Nonaka, Y. 2019. Hierarchical disentanglement of discriminative latent features for zero-shot learning. In CVPR, 11467–11476.

Verma, V. K.; Braham, D.; and Rai, P. 2020. Meta-Learning for Generalized Zero-Shot Learning. In AAAI, 6062–6069.

Wang, Q.; and Chen, K. 2017. Zero-shot visual recognition via bidirectional latent embedding. International Journal of Computer Vision 124(3): 356–383.

Wang, W.; Pu, Y.; Verma, V. K.; Fan, K.; Zhang, Y.; Chen, C.; Rai, P.; and Carin, L. 2018. Zero-shot learning via class-conditioned deep generative models. In AAAI, 4211–4218.

Welinder, P.; Branson, S.; Mita, T.; Wah, C.; Schroff, F.; Belongie, S.; and Perona, P. 2010. Caltech-UCSD birds 200.

Xian, Y.; Akata, Z.; Sharma, G.; Nguyen, Q.; Hein, M.; and Schiele, B. 2016. Latent embeddings for zero-shot classification. In CVPR, 69–77.

Xian, Y.; Lampert, C. H.; Schiele, B.; and Akata, Z. 2018a. Zero-shot learning: A comprehensive evaluation of the good, the bad and the ugly. IEEE Transactions on Pattern Analysis and Machine Intelligence 41(9): 2251–2265.

Xian, Y.; Lorenz, T.; Schiele, B.; and Akata, Z. 2018b. Feature generating networks for zero-shot learning. In CVPR, 5542–5551.

Yu, Y.; Ji, Z.; Han, J.; and Zhang, Z. 2020. Episode-Based Prototype Generating Network for Zero-Shot Learning. In CVPR, 14035–14044.

Zhang, L.; Xiang, T.; and Gong, S. 2017. Learning a deep embedding model for zero-shot learning. In CVPR, 2021–2030.

Zheng, Z.; and Saligrama, V. 2015. Zero-shot learning via semantic similarity embedding. In ICCV, 4166–4174.

Zhu, Y.; Elhoseiny, M.; Liu, B.; and Elgammal, A. 2018. A generative adversarial approach for zero-shot learning from noisy texts. In CVPR, 1004–1013.

Zhu, Y.; Xie, J.; Liu, B.; and Elgammal, A. 2019. Learning feature-to-feature translator by alternating back-propagation for generative zero-shot learning. In ICCV, 9844–9854.