Generalized Zero-Shot Learning via Disentangled Representation

Xiangyu Li*, Zhe Xu*, Kun Wei, Cheng Deng †
School of Electronic Engineering, Xidian University, Xi'an 710071, China
{xdu_xyLi, zhexu}@stu.xidian.edu.cn, {weikunsk, chdeng.xd}@gmail.com

Abstract

Zero-Shot Learning (ZSL) aims to recognize images belonging to unseen classes that are unavailable in the training process, while Generalized Zero-Shot Learning (GZSL) is a more realistic variant that both seen and unseen classes appear during testing. Most GZSL approaches achieve knowledge transfer based on the features of samples that inevitably contain information irrelevant to recognition, bringing negative influence for the performance. In this work, we propose a novel method, dubbed Disentangled-VAE, which aims to disentangle category-distilling factors and category-dispersing factors from visual as well as semantic features, respectively. In addition, a batch re-combining strategy on latent features is introduced to guide the disentanglement, encouraging the distilling latent features to be more discriminative for recognition. Extensive experiments demonstrate that our method outperforms the state-of-the-art approaches on four challenging benchmark datasets.

Introduction

Benefiting from the fast development of deep learning (LeCun, Bengio, and Hinton 2015; Ju et al. 2020b,a; Yang et al. 2020), supervised image classification (Krizhevsky, Sutskever, and Hinton 2012; He et al. 2016; Chang et al. 2020) has achieved remarkable success. However, the success is principally owed to abundant labeled data, which is costly or even impossible to acquire in most cases. Thus, Zero-Shot Learning (ZSL) (Palatucci et al. 2009; Larochelle, Erhan, and Bengio 2008) is proposed to tackle the problem above. In ZSL setting, the classes covered by the training images are referred to as the seen classes, while other classes are referred to as unseen classes, whose images are not available during training. Compared to ZSL that only unseen classes appear in the testing phase, Generalized Zero-Shot Learning (GZSL) is a more realistic and challenging variant of ZSL, that is, testing images can come from both seen and unseen classes.

Due to its promising application, GZSL has received extensive attention. Methods for GZSL can be roughly divided into three categories based on the space where the classification is conducted: (1) semantic-space (Lampert, Nickisch, and Harmeling 2009; Frome et al. 2013), (2) visual-space (Kumar Verma et al. 2018; Xian et al. 2018b; Felix et al. 2018), and (3) common-space (Schonfeld et al. 2019). Semantic-space based methods learn a mapping from visual space to semantic space on seen classes and generalize the learned mapping to unseen classes. Visual-space based methods formalize GZSL as a missing data problem and generate unseen images or image features to augment data. Common-space based methods embed visual features and semantic features into a common latent space.

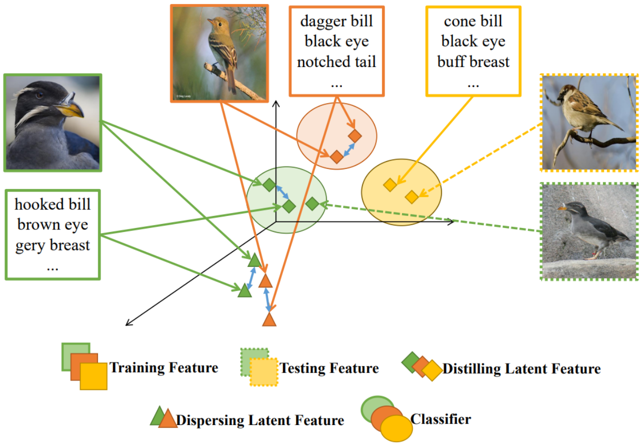


Figure 1: Illustration of the overall concept of our method. Distilling latent factors (in terms of rhombuses) and dispersing latent factors (in terms of triangles) are disentangled from paired visual and semantic features of seen classes. Classifier (in terms of ovalities) is trained using only the distilling latent features of both seen and semantic-part of unseen classes.

Most GZSL approaches (Xian, Schiele, and Akata 2017; Kumar Verma et al. 2018; Xian et al. 2018b; Felix et al. 2018; Schonfeld et al. 2019; Wei et al. 2019; Wei, Deng, and Yang 2020; Zhao et al. 2018; Min et al. 2020) achieve knowledge transfer based on overall representations of samples. The visual and semantic features inevitably contain information that is irrelevant to classification and influences

*Equal contribution.

†Corresponding author.

the classification results in a negative way. Concretely, visual features are extracted from pre-trained image classification model, e.g., ResNet (He et al. 2016), obtained from the entire image, while semantic features are per-class attributes or sentence embeddings extracted from sentences annotated per image averaged per class (Reed et al. 2016). Thus, a simple yet effective solution to improve the classification performance is to decouple the irrelevant features from the discriminative features.

Figure 1 illustrates the motivation of our method. In order to exclude the classification-irrelevant information, we present Disentangled-VAE to disentangle category-distilling factors and category-dispersing factors from visual features as well as semantic features respectively.

To be specific, category-distilling factors, corresponding to the discriminative part of features, are used for reconstruction as well as classification, while category-dispersing factors contain more irrelevant information for classification and are only used for reconstruction. In addition, considering that the category-dispersing factors disentangled from different samples contain few discriminative information, we frame a batch re-combining strategy on latent features to guide the disentanglement. Specifically, we shuffle the category-dispersing latent features in a batch and recombine them with the category-distilling ones. A classification loss is employed to maintain the category discriminability of the recombined latent features.

Our key contributions can be summarized as follows:

- We propose Disentangled-VAE for GZSL to disentangle category-distilling factors and category-dispersing factors from visual as well as semantic features, respectively. To the best of our knowledge, this is the first attempt to solve GZSL problem using disentangled representation learning.
- We introduce a batch re-combining strategy on latent features, which guides the disentanglement to obtain the category-distilling features for more accurate recognition.
- Extensive experimental results show that our method outperforms the state-of-the-art methods on four benchmark datasets.

Related Work

Zero-Shot Learning

ZSL has the ability to transfer knowledge to solve the problem of image classification even if the testing categories are not incorporated with training set. In ZSL, the instances of unseen classes are mapped into visual space only with semantic descriptions. However, the practical application of conventional ZSL methods is very poor due to mechanical setting that test samples are only sampled from the unseen classes. To address the disadvantages of ZSL, GZSL is more available since it not only learns information which can be adapted to unseen classes but also apply to the testing data from seen classes.

Inchoate ZSL methods usually established the rigid correspondence between original visual features and primitive label such as DAP (Lampert, Nickisch, and Harmeling 2009) and IAP (Lampert, Nickisch, and Harmeling 2013). Subsequently, more and more approaches aim to learn a mapping function to project from visual space to semantic space by each class attributes (Romera-Paredes and Torr 2015; Socher et al. 2013), or from semantic space to visual space (Zhang, Xiang, and Gong 2017). In addition to the model learning one-to-one mapping, other approaches also map visual and semantic features into a common space (Akata et al. 2015b; Sung et al. 2018).

One major drawback of all the methods mentioned above is that the training phase is conducted under the data of the seen classes, which leads to the over-fitting of seen class samples in the test phase even if samples have not been shown in the training phase in the GZSL setting. After that, generative methods can generate seen class and unseen class samples so that the prediction will not be biased towards the seen classes. According to category description or attribute, Generative Adversarial Networks (GAN) (Goodfellow et al. 2014) and Variational Auto-Encoder (VAE) (Kingma and Welling 2013) are used to generate visual features that approximate the original data, and the generated visual features and original features were combined to train the classifier to reduce the deviation problem of the classifier (Xian et al. 2018b). Felix et al. set L2 normalization via calculating Euclidean distance to constrain the features of reconstruction to ensure the high quality of the generated samples (Felix et al. 2018). Not only can GAN be used to generate unseen classes samples, but also VAE has the ability to solve the problem of inadequate unseen classes samples. Schonfeld et al. utilized double-deck VAE structure to reconstruct the visual and semantic features respectively, and used cross-modal reconstruction to establish the deep relationship between visual and semantic space to align them (Schonfeld et al. 2019).

Nonetheless, none of the methods mentioned above take into account the fact that the features of each category are redundant in the classification process, which is called category redundancy. Category redundancy information exists not only in visual space, but also in semantic space, which only influences the classification results. Therefore, we propose Disentangled-VAE model to extract category-distilling features which is sensitive for classification.

Disentangled Representation

With the advance of deep generative models, many efforts have been made on disentangled representations. Higgins et al. set a heavier weight on KL divergence for better disentanglement (Higgins et al. 2016). Kim and Mnih derived a Total Correlation (TC) from the KL divergence and emphasized this TC term as the key point in disentangled representation learning (Kim and Mnih 2018). Tran, Yin, and Liu proposed a framework to separate the information of posture and character identity for face recognition (Tran, Yin, and Liu 2017). Jiang et al. disentangled the content information and style information of images in order to generate the style transfered images (Jiang et al. 2020). Zhu et al. proposed a Self-Supervised Sequential VAE model which was use to disentangle the time-varying variables and time-invariant variables of video and audio sequences (Zhu et al.

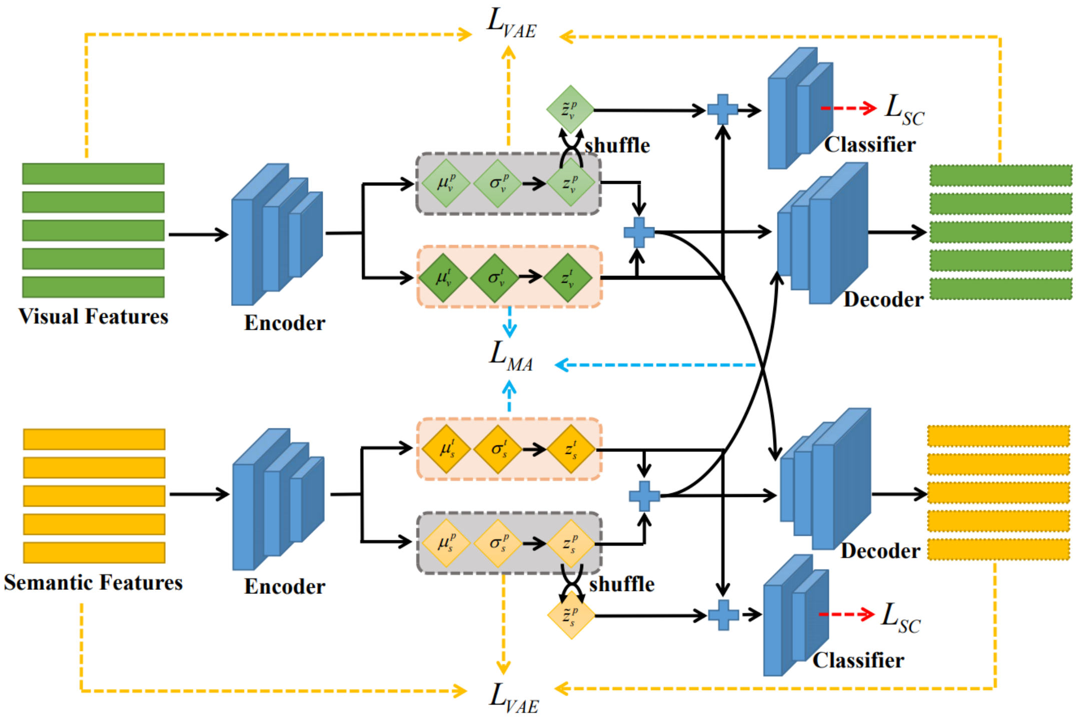


Figure 2: The framework of our Disentangled-VAE method. The proposed model consists of two parallel VAEs with two branches each: one for visual features, the other for semantic features. Distilling latent factors and dispersing latent factors are disentangled from visual as well as semantic features and only distilling ones will be used for classification. We accomplish disentanglement based on three kinds of loss functions, namely: (1) Variational Auto-Encoder loss $\mathcal{L}_{VAE}$, (2) Shuffling classification loss $\mathcal{L}_{SC}$, and (3) Modality alignment loss $\mathcal{L}_{MA}$.

2020). Meanwhile, more and more image-to-image translation models (Almahairi et al. 2018; Huang et al. 2018; Lee et al. 2018) also used the idea of decoupling, which separate the original features into domain-invariant content features and domain-specific attribute vectors to improve the performance.

Based on the effectiveness of the methods mentioned above, we propose a framework which consists of double-deck VAE structure for disentangling the visual and semantic features so that the input features can be more discriminative in the process of classification.

Methodology

In this section, we first present the definition of GZSL. Then we briefly introduce VAE, the basic building block of our model. Next the proposed method is explained in detail. Finally, the training and inference process of our method is summarized at the end of the section.

Problem Definition

Denote $X \subseteq \mathbb{R}^{d_1}$ as the visual space and $C \subseteq \mathbb{R}^{d_2}$ as the semantic space. $Y^S = \{y_i^s | i = 1, 2, \ldots, N_s\}$ and $Y^U = \{y_j^u | j = 1, 2, \ldots, N_u\}$ is referred to as the set of seen categories and the set of unseen categories. In addition, seen and unseen categories are disjoint, i.e., $Y^S \cap Y^U = \emptyset$.

Given training examples $\{(x, y^s, c) | x \in X, y^s \in Y^S, c \in C\}$ of seen classes and auxiliary data $\{(y^u, c) | y^u \in Y^U, c \in C\}$ of unseen classes, ZSL aims to learn a classifier $f^{ZSL}: X \to Y^U$ that can recognize a testing instance $x \in X$ belonging to the unseen classes whose instances are not available during training. GZSL is a more realistic and challenging variant of ZSL, aiming to learn a classifier $f^{GZSL}: X \to Y^S \cup Y^U$.

Variational Auto-Encoder

Auto-Encoder (AE) learns the latent representation of input by minimizing reconstruction error while VAE is the variational counterpart of AE. Assuming a specific prior $p(z)$ on the latent space and parameterizing $p(x|z)$ as well as $q(z|x)$ with deep neural networks, one can get parameters ϕ and θ by optimizing objective below:

$$\max_{\phi,\theta} \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p(z)), \quad (1)$$

where the first term can be seen as the reconstruction error, similar to classical AE. The second term is the KL-divergence between distributions $q(z|x)$ and $p(z)$, constraining the encoder distribution to match the factorized prior, e.g., Gaussian distribution.

In order to obtain a differentiable estimator of the variational lower bound, a trick called reparameterization is used:

$$z = \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon, \quad (2)$$

where $\mu_\phi(x)$ and $\sigma_\phi(x)$ are the outputs of encoders, representing the mean and variance of the posterior distributions. $\odot$ is the element-wise product and $\epsilon \sim \mathcal{N}(0, 1)$ is an auxiliary noise variable.

Disentangled-VAE

Just like any other machine learning task, the performance of GZSL is heavily dependent on the data representation. However, representations of most exsiting GZSL approaches are evolved from overall features of instances, extracted from pre-trained model or hand-annotated. These features inevitably contain information irrelevant to classification and eventually bring negative influence to classification. Therefore, designing models that can exclude irrelevant information to mitigate its negative impact is a keypoint to improve the classification performance of GZSL.

In order to exclude the irrelevant information, we present Disentangled-VAE to disentangle category-distilling factors and category-dispersing factors from visual features as well as semantic features, as illustrated in Figure 2. Differing from vanilla VAE that parameterize encoder distribution $q(z|x)$ using neural network, our objective is to learn the encoder distribution $q(z^t, z^p|x)$, where z^t and z^p represent category-distilling factors and category-dispersing factors, respectively.

To this end, we build a model with two parallel VAEs and each with two branches. The visual encoder E_v and semantic encoder E_s map the input visual features x and semantic features c to the corresponding ditilling latent features z_v^t, z_s^t and dispersing latent features z_v^p, z_s^p, respectively. In addition, the visual decoder D_v and semantic decoder D_s are employed for construction and classifiers F_v as well as F_s are utilized for better disentanglement.

Variational Auto-Encoder loss. Our model consists of two VAEs: one for visual feature and the other for semantic feature. The Variational Auto-Encoder loss constrains the encoder distributions of distilling branch and dispersing branch to match the factorized priors, respectively. In the meantime, it guarantees that the input features can be reconstructed from distilling latent features and dispersing latent features by minimizing reconstruction error.

Denote encoder distrubutions of visual modality as $q_\phi(z_v^t|x)$ and $q_\phi(z_v^p|x)$, belonging to distilling branch and dispersing branch respectively. Denote decoder distribution of visual modality as $p_\theta(x|z_v^t, z_v^p)$. The priors $p(z_v^t)$ and $p(z_v^p)$ are both standard Gaussian distributions. Loss function of visual VAE is:

$$\begin{aligned}\mathcal{L}_{VAE}^v = &-\mathbb{E}_{q_\phi(z_v^t, z_v^p|x)}[\log p_\theta(x|z_v^t, z_v^p)] \\ &+ D_{KL}(q_\phi(z_v^t|x)||p(z_v^t)) \\ &+ \alpha D_{KL}(q_\phi(z_v^p|x)||p(z_v^p)),\end{aligned} \tag{3}$$

where the first term is reconstruction error. The second and third term are the KL-divergence between encoder distributions and priors for distilling and dispersing branch, respectively. α is the weighting factor. Noting that dispersing latent features are redundant for classification but indispensable for reconstruction, we use the sum of z_v^t and z_v^p for reconstruction.

Similarly, loss function of semantic VAE is:

$$\begin{aligned}\mathcal{L}_{VAE}^s = &-\mathbb{E}_{q_\phi(z_s^t, z_s^p|c)}[\log p_\theta(c|z_s^t, z_s^p)] \\ &+ D_{KL}(q_\phi(z_s^t|c)||p(z_s^t)) \\ &+ \alpha D_{KL}(q_\phi(z_s^p|c)||p(z_s^p)).\end{aligned} \tag{4}$$

The final VAE loss is combined as:

$$\mathcal{L}_{VAE} = \mathcal{L}_{VAE}^v + \mathcal{L}_{VAE}^s. \tag{5}$$

Shuffling classification loss. Considering that the different category-dispelling factors contain few discriminative information and do not change the classification results, we add two auxiliary classifiers to guide the disentanglement, each for a modality.

Denote $Z_m^T = \{z_{m,1}^t, z_{m,2}^t, \ldots, z_{m,N}^t\}$ and $Z_m^P = \{z_{m,1}^p, z_{m,2}^p, \ldots, z_{m,N}^p\}$ as sets of distilling and dispersing latent features in a batch, respectively. N is the batch size and $m \in \{v, s\}$ represents different modality. We shuffle Z_m^P and get $\tilde{Z}_m^P = \{\tilde{z}_{m,1}^p, \tilde{z}_{m,2}^p, \ldots, \tilde{z}_{m,N}^p\}$. Shuffling classification loss below is employed for better disentanglement:

$$\begin{aligned}\mathcal{L}_{SC} = \sum_m \sum_{i=1}^{N} &- F_m(z_{m,i}^t + z_{m,i}^p)[y_i] \\ &- F_m(z_{m,i}^t + \tilde{z}_{m,i}^p)[y_i],\end{aligned} \tag{6}$$

where the first and the second term is the negative log likelihood loss of original and re-combined latent features, respectively. F_m is the classifiers we add that output the log-likelihood of belonging to all classes. $y_i \in \{0, 1, \ldots, N-1\}$ is the label of i^{th} latent features in the batch and N is the number of classes.

Modality alignment loss. Aligning visual and semantic features in the latent space is of great importance for GZSL, which guarantees the generalization from seen to unseen classes. We employ cross-reconstruction loss and distribution-distance loss for modality alignment. The cross-reconstruction loss is denoted as :

$$\mathcal{L}_{CR} = |x - D_v(E_s(c))| + |c - D_s(E_v(x))|, \tag{7}$$

where the first term is the cross-reconstruction error from semantic modality to visual modality and the second term inversely. E_v, E_s and D_v, D_s are encoders and decoders of corresponding modality. Note that the outputs of our encoders consist of two parts, $E_s(c) = (z_s^t, z_s^p)$ for example, meaning that we reconstruct inputs based on both distilling and dispersing latent features.

The distribution-distance loss is:

$$\begin{aligned}\mathcal{L}_{DD} = &(||\mu_v^t - \mu_s^t||_2^2 + ||{\sigma_v^t}^{\frac{1}{2}} - {\sigma_s^t}^{\frac{1}{2}}||_{Frobenius}^2)^{\frac{1}{2}} \\ &+ (||\mu_v^p - \mu_s^p||_2^2 + ||{\sigma_v^p}^{\frac{1}{2}} - {\sigma_s^p}^{\frac{1}{2}}||_{Frobenius}^2)^{\frac{1}{2}},\end{aligned} \tag{8}$$

where the first term is the distribution distance between distilling branches of two modalities and the second term for dispersing branches. μ and σ are the mean and variance of the posterior distributions which are the outputs of encoders.

The final modality alignment loss is:

$$\mathcal{L}_{MA} = \mathcal{L}_{CR} + \beta \mathcal{L}_{DD}, \tag{9}$$

where β is the weighting coefficient of the distribution distance loss.

Overall loss function. As a summary, the overall loss for our Disentangled-VAE method is expressed as follows:

$$\mathcal{L} = \mathcal{L}_{VAE} + \gamma\mathcal{L}_{SC} + \lambda\mathcal{L}_{MA}, \tag{10}$$

where $\mathcal{L}_{VAE}$ is the basic VAE loss, $\mathcal{L}_{SC}$ is the shuffling classification loss to guide the disentanglement, and $\mathcal{L}_{MA}$ is the modality alignment loss to align the latent visual and semantic representation. γ and λ are the weighting coefficients of shuffling classification loss and modality alignment, respectively.

Training and Inference

Given visual features extracted from pre-trained model and semantic features hand-annotated , we solve GZSL problem in three steps: (1) Training Disentangled-VAE, (2) Training classifier, and (3) Inference.

Specifically, paired visual and semantic features of seen classes are employed to train the Disentangled-VAE model based on equation (10). Once trained, Disentangled-VAE is able to disentangle distilling and dispersing latent factors from features that belong to seen classes as well as unseen classes. Different from most GZSL methods that train classifiers based on overall features, our classifier is trained using only the distilling latent features of both seen and semantic-part of unseen classes. During inference, visual features x of images extracted from pre-trained model are firstly mapped to the distilling latent features z_t using the learned visual encoder E_v. The final classification results are obtained from z_t based on the learned classifier. Note that images can come from both unseen and seen classes.

Experiments

In this section, all datasets and evaluation protocol are introduced in detail. In addition, we present the the implementation details as well as the comparison of experimental results with other state-of-the-art methods. Eventually, ablation study proves the effectiveness our method.

Datasets and Evaluation Protocol

We evaluate our model on four popular datasets: Caltech-UCSD-Birds 200-2011 dataset (CUB) (Welinder et al. 2010), Animals with Attributes 1 (AWA1) (Lampert, Nickisch, and Harmeling 2009) and 2 (AWA2) (Xian et al. 2018a), SUN Attribute dataset (SUN) (Patterson and Hays 2012). The CUB dataset contains a total of 200 bird species, 150 of which are seen and 50 of which are unseen. Since there are subtle differences between the categories of birds, it is necessary for the learning model to extract more discriminative features. AWA1 and AWA2 are datasets which are commonly used for animal classification, consisting of 40 seen classes and 10 unseen classes. SUN is a large scenario-style dataset with 645 seen classes and 72 unseen classes. The detailed information of each dataset is summarized in Table 1. To avoid violating the zero-shot setting, we adopt the typical training splits proposed by (Xian et al. 2018a) for training split so that test samples can be disjoint from training samples which ResNet-101 is trained with on each dataset.

Dataset	d_1	d_2	N_s	N_u	X_a
CUB	2048	312	150	50	11788
AWA1	2048	85	40	10	40475
AWA2	2048	85	40	10	37322
SUN	2048	102	645	72	15339

Table 1: Datasets used in our experiments and their statistics, in terms of dimensionality of visual features d_1, dimensionality of semantic features d_2, number of seen classes N_s, number of unseen classes N_u and number of all instances X_a.

In addition to datasets setting, the evaluation protocol is shown as following:

- U : the average accuracy of per-class on test images from unseen classes, which represents the capacity of classifying unseen classes samples.
- S : the average accuracy of per-class on test images from seen classes, which is used to represent the capacity of classifying seen classes samples.
- H : the harmonic mean value, which is formulated as

$$H = \frac{2 \times U \times S}{U + S}. \tag{11}$$

Implementation Details

Following the setting in other methods (Xian et al. 2018b; Schonfeld et al. 2019), we utilize a pre-trained ResNet-101 to extract visual features which are represented as 2048-dimensional vectors. Semantic features are per-class attributes annotated by humen. The encoder E_v and E_s, decoder D_v and D_s consist of multilayer perceptron (MLP) with two hidden layers. We utilize 1560, 1660 hidden units for encoder E_v and 1450, 660 hidden units for encoder E_s. The size of latent feature is implemented as 64 in the whole datasets. Due to final classification being trained in the latent space, we set the same size as latent feature for classifier. Our approach is implemented with PyTorch(Paszke et al. 2019) and optimized by ADAM optimizer (Kingma and Ba 2014). In addition, we set learning rate as 0.00015, batch size as 50 and epochs as 150. After the process of training VAE model is complete, the final classifier will be trained with category-distilling variables. Due to the difference between each dataset, the proportion of seen classes samples and unseen classes samples is different and inspired the results of many experiments we have made, we use a fixed dataset with 200 samples per seen class and 400 samples per unseen class in CUB dataset, 200, 460 in AWA1 dataset, 200, 480 in AWA2 dataset, and 200, 410 in SUN dataset.

Comparison with State-of-the-Art Methods

Baseline models We compare our model with fifteen state-of-the-art methods. Among them, the typical GZSL methods ALE (Akata et al. 2015a), DeViSE (Frome et al. 2013), ReViSE (Hubert Tsai, Huang, and Salakhutdinov

Model	CUB			AWA1			AWA2			SUN		
	U	S	H	U	S	H	U	S	H	U	S	H
ALE	23.7	62.8	34.4	16.8	76.1	27.5	14.0	81.8	23.9	21.8	33.1	26.3
DeViSE	23.8	53.0	32.8	13.4	68.7	22.4	17.1	74.7	27.8	16.9	27.4	20.9
ReViSE	37.6	28.3	32.8	46.1	37.1	41.1	46.4	39.7	42.8	24.3	20.1	22.0
DEM	19.6	57.9	29.2	32.8	84.7	47.3	30.5	86.4	45.1	19.6	57.9	29.2
SP-AEN	34.7	70.6	46.6	-	-	-	23.3	90.9	37.1	24.9	38.6	30.3
cycle-CLSWGAN	59.3	47.9	53.0	63.4	59.6	59.8	-	-	-	33.8	47.2	39.4
f-CLSWGAN	57.7	43.7	49.7	61.4	57.9	59.6	53.8	68.2	60.2	36.6	42.6	39.4
LiGAN	46.5	57.9	51.6	52.6	76.3	62.3	54.3	68.5	60.6	42.9	37.8	40.2
f-VAEGAN-D2	60.1	48.4	53.6	70.6	57.6	63.5	-	-	-	38.0	45.1	41.3
CADA-VAE	51.6	53.5	52.4	57.3	72.8	64.1	55.8	75.0	63.9	47.2	35.7	40.6
DASCN	59.0	45.9	51.6	68.0	59.3	63.4	-	-	-	38.5	42.4	40.3
LsrGAN	59.1	48.1	53.0	74.6	54.6	63.0	74.6	54.6	63.0	37.7	44.8	40.9
OCD	59.9	44.8	51.3	-	-	-	73.4	59.5	65.7	42.9	44.8	**43.8**
E-PGN	57.2	48.5	52.5	86.3	52.6	65.3	83.6	48.0	61.0	-	-	-
DAZLE	42.0	65.3	51.1	-	-	-	25.7	82.5	39.2	25.7	82.5	25.8
Our model	51.1	58.2	**54.4**	60.7	72.9	**66.2**	56.9	80.2	**66.6**	36.6	47.6	41.4

Table 2: Performance of GZSL on four classification benchmarks. U and S are the recognition accuracies tested on seen and unseen classes, respectively. H is the harmonic mean of U and S in GZSL setting.(Top one performance is highlighted)

Model	CUB			AWA1			AWA2			SUN		
	U	S	H	U	S	H	U	S	H	U	S	H
Base model	51.6	53.5	52.4	57.3	72.8	64.1	55.8	75.0	63.9	47.2	35.7	40.6
+dis	50.0	58.0	53.7	57.6	73.0	65.0	56.1	78.9	65.6	46.4	36.9	41.1
+dis+re	51.1	58.2	**54.4**	60.7	72.9	**66.2**	56.9	80.2	**66.6**	36.6	47.6	**41.4**

Table 3: Ablation study on four datasets. dis and re represent the disentanglement and batch re-combining strategy, respectively.

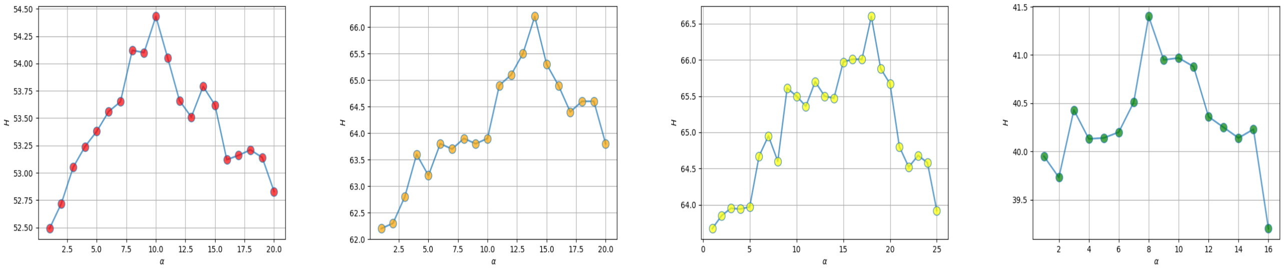
2017), DEM (Zhang, Xiang, and Gong 2017) and SP-AEN (Chen et al. 2018) aim to learn a linear or nonlinear mapping function to find the relation between semantic space and visual space. In addition to the idea of projection, the generative methods are designed to generate synthetic unseen class samples by using VAE or GAN to approximate the distribution of class visual samples as the function of class semantic descriptions such as f-CLSWGAN (Xian et al. 2018b), cycle-CLSWGAN (Felix et al. 2018), LiGAN (Li et al. 2019), f-VAEGAN-D2 (Xian et al. 2019), DASCN (Ni, Zhang, and Xie 2019), CADA-VAE (Schonfeld et al. 2019), OCD (Keshari, Singh, and Vatsa 2020), DAZLE (Huynh and Elhamifar 2020), E-PGN (Yu et al. 2020) and LsrGAN (Vyas, Venkateswara, and Panchanathan 2020).

Results Table 2 represents the results of the comparing approaches and our method, which significantly indicates that our proposed method is superior to the state-of-the-art methods mentioned in the table. Or rather, in our method, the value of H can reach 54.4% on CUB dataset, 66.2% on AWA1, 66.6% on AWA2, and 41.4% on SUN. Specifically compared with the original CADA-VAE model, our method increases the H value of our model from 52.4% to 54.4% on CUB dataset, from 64.1% to 66.2% on AWA1, from 63.9% to 66.6% on AWA2, and from 40.6% to 41.4% on SUN. The reason why our method performs better is that our model separates the discriminative features of classification before inputting to the classifier, which enables the classifier to learn the difference between categories better. The performance boost is attributed to the effectiveness of our model which transfers the knowledge from seen classes and excavates category-distilling visual features of the unseen classes.

Ablation Study

Effectiveness of disentanglement and batch re-combining strategy. In order to prove that the effectiveness of our model depends on the disentanglement of the features and the batch re-combining strategy. We make an ablation study to verify the importance of these two componets separately. The results of our model with different modules are presented in Table 3.

By ablation experiments results, we can see that after adding the disentanglement we proposed to the basic model, compared with the basic model, the H value has increased from 52.4% to 53.7% on CUB dataset, and from 64.1% to 65.0% on AWA1, from 63.9% to 66.6% on AWA2, from 40.6% to 41.1% on SUN, which shows that the model can extract features with better discrimination, which is called category-distilling feature, for classification after the disentanglement operation. After adding the batch re-combining strategy we proposed, our model has improved from 53.7% to 54.4% on CUB dataset, from 65.0% to 66.2% on AWA1, and from 65.6% to 66.6% on AWA2, and on SUN the episode has increased from 41.1% to 41.4%. This phenomenon shows that the batch re-combining strategy as an



(a) CUB (b) AWA1 (c) AWA2 (d) SUN

Figure 3: The influence of the weighting coefficient α. We measure the harmonic mean accuracy (H) on CUB, AWA1, AWA2, and SUN.

Latent Features	CUB			AWA1			AWA2			SUN		
	U	S	H	U	S	H	U	S	H	U	S	H
z^p	43.4	53.1	47.7	44.9	68.2	54.1	42.3	68.8	52.4	35.7	30.5	32.9
$z^p + z^t$	49.6	56.8	52.9	62.4	65.8	64.1	54.8	78.8	64.8	42.5	36.0	39.0
z^t	51.1	58.2	**54.4**	60.7	72.9	**66.2**	56.9	80.2	**66.6**	36.6	47.6	**41.4**

Table 4: Discriminability of different latent features on CUB, AWA1, AWA2, and SUN. z^p and z^t represent the dispersing and distilling latent features, respectively. U, S, and H are the average accuracy of unseen classes, the average accuracy of seen classes and the harmonic mean accuracy, respectively.

auxiliary operation can effectively guide disentanglement to extract better category-distilling features.

Choice of weighting coefficient α. α is the weighting coefficient of the KL-divergence for the dispersing branch, same for both visual VAE and semantic VAE. When α increases, the encoder distributions of dispersing branch will be forced to better match the factorized unit Gaussian priors and the weight of reconstruction that the dispersing branch contributes to will decline correspondingly.

In this experiment, we vary α from 1 to 20 on CUB, from 1 to 20 on AWA1, from 1 to 25 on AWA2, and from 1 to 16 on SUN. The harmonic mean accuracies increase until they achieve their peak accuracies at $\alpha = 10, 14, 18, 8$ for different datasets, as shown in Figure 3. We conclude from this experiment that a tradeoff between reconstruction error and KL-divergence of dispersing branch is gained when $\alpha = 10, 14, 18, 8$ on CUB, AWA1, AWA2, and SUN, respectively. In fact, the main difference of the constraint on different latent features is the weighting coefficient except the batch re-combining strategy. As shown in the experiment, optimal performance is achieved when the weighting coefficient of dispersing branch α is greater than one and the weighting coefficient of distilling branch is set to one. We can also conclude that dipersing latent features better match the factorized prior than the distilling ones, while distilling latent features contribute more to reconstruction than dispersing ones.

Discriminability of different latent features. To investigate what is encoded in distilling and dispersing latent features, we add classifiers on the learned latent features to obtain the discriminability of different features. Intuitively, the distilling latent features are supposed to be more discriminative than the dispersing ones. We report the accuracy for different latent features on CUB, AWA1, AWA2, and SUN.

We can obviously observe in Table 4 that the dispersing latent features are less discriminative as expected and the accuracy increases with the addition of the distilling latent features, achieving almost the same results as CADA-VAE where similar overall latent features are used for classification. The distilling latent features are proved to be the best latent features for classification, whose discriminability surpass others by a large margin. We can conclude from this experiment that the dispersing latent features are less discriminative due to the inclusion of irrelevant information and our method can successfully disentangle category-ditilling latent features from the overall features for more accurate classification.

Conclusion

In this work, we propose Disentangled-VAE to excavate category-distilling information for Generalized Zero-Shot Learning. The proposed model can separate the latent features into category-distilling features for classification as well as reconstruction and category-dispersing features only for reconstruction. In addition, we design a batch recombining strategy as an auxiliary operation for better disentanglement. The training process provides us with an encoder to encode discriminative features for classification from testing features in the latent space, where a linear softmax classifier can be trained to easily recognize different categories. We further prove the necessity of the auxiliary operation through ablation experiments and verify the effectiveness of our method on four benchmark datasets.

Acknowledgments

Our work was supported in part by the National Natural Science Foundation of China under Grant 62071361, the National Key R&D Program of China under Grant 2017YFE0104100, and the China Research Project under Grant 6141B07270429.

References

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2927–2936.

Almahairi, A.; Rajeswar, S.; Sordoni, A.; Bachman, P.; and Courville, A. 2018. Augmented cyclegan: Learning many-to-many mappings from unpaired data. *arXiv preprint arXiv:1802.10151* .

Chang, D.; Ding, Y.; Xie, J.; Bhunia, A. K.; Li, X.; Ma, Z.; Wu, M.; Guo, J.; and Song, Y. Z. 2020. The Devil is in the Channels: Mutual-Channel Loss for Fine-Grained Image Classification. *IEEE Transactions on Image Processing* 29: 4683–4695. doi:10.1109/TIP.2020.2973812.

Chen, L.; Zhang, H.; Xiao, J.; Liu, W.; and Chang, S.-F. 2018. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1043–1052.

Felix, R.; Kumar, V. B.; Reid, I.; and Carneiro, G. 2018. Multi-modal cycle-consistent generalized zero-shot learning. In *Proceedings of the European Conference on Computer Vision (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 *Advances in neural information processing systems*, 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 *Advances in neural information processing systems*, 2672–2680.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2016. beta-vae: Learning basic visual concepts with a constrained variational framework .

Huang, X.; Liu, M.-Y.; Belongie, S.; and Kautz, J. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 172–189.

Hubert Tsai, Y.-H.; Huang, L.-K.; and Salakhutdinov, R. 2017. Learning robust visual-semantic embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, 3571–3580.

Huynh, D.; and Elhamifar, E. 2020. Fine-Grained Generalized Zero-Shot Learning via Dense Attribute-Based Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4483–4493.

Jiang, W.; Liu, S.; Gao, C.; Cao, J.; He, R.; Feng, J.; and Yan, S. 2020. Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5194–5202.

Ju, Y.; Dong, X.; Wang, Y.; Qi, L.; and Dong, J. 2020a. A dual-cue network for multispectral photometric stereo. *Pattern Recognition* 100: 107162.

Ju, Y.; Lam, K.-M.; Chen, Y.; Qi, L.; and Dong, J. 2020b. Pay Attention to Devils: A Photometric Stereo Network for Better Details. In Bessiere, C., ed., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, 694–700. International Joint Conferences on Artificial Intelligence Organization. Main track.

Keshari, R.; Singh, R.; and Vatsa, M. 2020. Generalized Zero-Shot Learning Via Over-Complete Distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13300–13308.

Kim, H.; and Mnih, A. 2018. Disentangling by factorising. *arXiv preprint arXiv:1802.05983* .

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

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

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Kumar Verma, V.; Arora, G.; Mishra, A.; and Rai, P. 2018. Generalized zero-shot learning via synthesized examples. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4281–4289.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 951–958. IEEE.

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.

Larochelle, H.; Erhan, D.; and Bengio, Y. 2008. Zero-data learning of new tasks. In *AAAI*, volume 1, 3.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *nature* 521(7553): 436–444.

Lee, H.-Y.; Tseng, H.-Y.; Huang, J.-B.; Singh, M.; and Yang, M.-H. 2018. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, 35–51.

Li, J.; Jing, M.; Lu, K.; Ding, Z.; Zhu, L.; and Huang, Z. 2019. Leveraging the invariant side of generative zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7402–7411.

Min, S.; Yao, H.; Xie, H.; Wang, C.; Zha, Z.-J.; and Zhang, Y. 2020. Domain-aware Visual Bias Eliminating for Generalized Zero-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12664–12673.

Ni, J.; Zhang, S.; and Xie, H. 2019. Dual adversarial semantics-consistent network for generalized zero-shot learning. In *Advances in Neural Information Processing Systems*, 6146–6157.

Palatucci, M.; Pomerleau, D.; Hinton, G. E.; and Mitchell, T. M. 2009. Zero-shot learning with semantic output codes. In *Advances in neural information processing systems*, 1410–1418.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, 8026–8037.

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

Reed, S.; Akata, Z.; Lee, H.; and Schiele, B. 2016. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 49–58.

Romera-Paredes, B.; and Torr, P. 2015. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8247–8255.

Socher, R.; Ganjoo, M.; Manning, C. D.; and Ng, A. 2013. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, 935–943.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1199–1208.

Tran, L.; Yin, X.; and Liu, X. 2017. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1415–1424.

Vyas, M. R.; Venkateswara, H.; and Panchanathan, S. 2020. Leveraging Seen and Unseen Semantic Relationships for Generative Zero-Shot Learning. *arXiv preprint arXiv:2007.09549* .

Wei, K.; Deng, C.; and Yang, X. 2020. Lifelong Zero-Shot Learning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*. International Joint Conferences on Artificial Intelligence Organization.

Wei, K.; Yang, M.; Wang, H.; Deng, C.; and Liu, X. 2019. Adversarial Fine-Grained Composition Learning for Unseen Attribute-Object Recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 3741–3749.

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

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5542–5551.

Xian, Y.; Schiele, B.; and Akata, Z. 2017. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4582–4591.

Xian, Y.; Sharma, S.; Schiele, B.; and Akata, Z. 2019. f-VAEGAN-D2: A feature generating framework for any-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10275–10284.

Yang, X.; Deng, C.; Wei, K.; Yan, J.; and Liu, W. 2020. Adversarial Learning for Robust Deep Clustering. *Advances in Neural Information Processing Systems* 33.

Yu, Y.; Ji, Z.; Han, J.; and Zhang, Z. 2020. Episode-Based Prototype Generating Network for Zero-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14035–14044.

Zhang, L.; Xiang, T.; and Gong, S. 2017. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021–2030.

Zhao, B.; Sun, X.; Fu, Y.; Yao, Y.; and Wang, Y. 2018. Msplit lbi: Realizing feature selection and dense estimation simultaneously in few-shot and zero-shot learning. *ICML* .

Zhu, Y.; Min, M. R.; Kadav, A.; and Graf, H. P. 2020. S3VAE: Self-Supervised Sequential VAE for Representation Disentanglement and Data Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6538–6547.