Hierarchical Disentanglement of Discriminative Latent Features for Zero-shot Learning

Bin Tong * † Chao Wang * † Martin Klinkigt †
Yoshiyuki Kobayashi † Yuuichi Nonaka †
† R&D Group, Hitachi, Ltd., Japan
‡ Ocean University of China

{bin.tong.hh, martin.klinkigt.ut}@hitachi.com, chaowangplus@gmail.com
{yoshiyuki.kobayashi.gp, yuichi.nonaka.zy}@hitachi.com

Abstract

Most studies in zero-shot learning model the relationship, in the form of a classifier or mapping, between features from images of seen classes and their attributes. Therefore, the degree of a model’s generalization ability for recognizing unseen images is highly constrained by that of image features and attributes. In this paper, we discuss two questions about generalization that are seldom discussed. Are image features trained with samples of seen classes expressive enough to capture the discriminative information for both seen and unseen classes? Is the relationship learned from seen image features and attributes sufficiently generalized to recognize unseen classes? To answer these two questions, we propose a model to learn discriminative and generalizable representations from image features under an auto-encoder framework. The discriminative latent features are learned through a group-wise disentanglement over feature groups with a hierarchical structure. On popular benchmark data sets, a significant improvement over state-of-the-art methods in tasks of typical and generalized zero-shot learning verifies the generalization ability of latent features for recognizing unseen images.

1. Introduction

The significant performance improvement of deep neural networks in recent years is partly due to the wide availability of large labeled datasets. However, for some uncommon objects, only a limited number of samples can be provided, and new categories of objects may even emerge dynamically. In such a situation, problems may arise regarding state-of-the-art methods recognizing new categories. Zero-shot learning [27, 45, 47, 15, 32, 16, 7, 6, 48] addresses the problem of recognizing objects of unseen classes by transferring knowledge from seen classes via mid-level descriptors, such as attributes [27], which bridge semantically low-level image features and high-level concepts such as class labels. For transferring such knowledge, most studies in zero-shot learning model the relationship, in the form of a mapping or classifier, between visual features and attributes. However, this relationship has a limited capability for recognizing unseen images, and it even performs worse for generalized zero-shot learning, in which both seen and unseen classes have to be recognized. We argue that this incapability boils down to the following limitations that are however seldom discussed in previous studies.

For convolutional neural network (CNN) features trained with seen images, seen classes are well separated in the feature space. However, these features are less capable of discriminating seen and unseen classes, which may hurt the performance of generalized zero-shot learning. Figure 1(a) shows that seen classes unnecessarily overlap unseen classes if image features are trained with classifying seen classes only. Figure 1(b) shows that the degree of discrimination in both seen and unseen classes significantly decreases, compared to that in seen classes. Figure 1(c) shows how unseen classes separate from seen classes when the dimensions of features are ranked by variances in descending order. We observed that not all dimensions of features contribute to the discrimination between seen and unseen classes. Features with low variances even hurt the discrimination between seen and unseen classes. The above observations explain part of the reason why state-of-the-art methods behave poorly in recognizing both seen and unseen images in more realistic settings such as generalized zero-shot learning.

*contribute equally to this paper. Chao Wang did this work during the internship at Hitachi.
Figure 1. Observations of image features. For image features, we use Res-net features [19] for data sets, such as aPY, CUB, SUN, AWA1 and AWA2. Figure 1(a) is t-SNE visualization of features in AWA1. ‘s’ and ‘u’ denote seen and unseen classes with orange and green colors, respectively. In both seen and unseen, different classes correspond to different shapes and shades of the same color. Other data sets behave similarly. Figure 1(b) and Figure 1(c) show the degree of discrimination between colors, respectively. In both seen and unseen, different classes correspond to different shapes and shades of the same color. Other data sets ranked in descending order and columns $x_i$ are ranked accordingly.

For most methods in zero-shot learning, attributes and image features are mapped to a common visual-semantic space. In inference phase, an unseen image is categorized with the label of the closest attribute representation via nearest-neighbor search. In this framework, correct recognition heavily depends on how well the features of seen classes and human-made attributes can encode the information invariant to both seen and unseen classes. As explained above, image features trained with seen classes are insufficient to encode all discrimination information in both seen and unseen classes. In addition, not all attributes are visually discriminative, as they are not designed for classification tasks. The mapping learned from seen classes can not encode all possible combinations of visual appearance and attributes only appearing in unseen classes. Therefore, some mappings of unseen images are more likely to fall into areas of seen images’ attributes rather than those of unseen images’ attributes [14].

The above limitations suggest that both dimensions of image features and hand-made attributes are highly correlated, such that the mapping learned from seen classes is variant or sensitive to unseen classes. Unlike previous studies, we address the zero-shot learning from a different perspective. We learn discriminative latent representations from image features, which are invariant to unseen classes. Disentangled representation is one choice of such latent representations. Bengio et al. described it in [5]: a representation where a change in one dimension corresponds to a change in one factor of variation, while being relatively invariant to changes in other factors. Disentangled representation can generalize the knowledge beyond the training distribution by recombining previously-learned independent factors [5, 9, 20].

In this work, we explain the image feature space from perspectives of discrimination and interpretability. We factorize a CNN feature into three latent features, including semantic feature, non-semantic but discriminative feature, and non-discriminative feature. Unlike previous studies that have dimension-wise disentanglement of features, we perform hierarchical disentanglement over groups of dimensions. The dimensions in a group can be correlated each other but in whole represent a most fine-grained concept of an image. The fusion of semantic and non-semantic latent features can be taken as a variant of an image feature, which is discriminative but more generalizable than the original image feature. This variant of an image feature and attributes are further used for learning a visual-semantic mapping. Furthermore, learning such a mapping can be jointly trained with disentangling the latent features, which creates a unified framework for zero-shot recognition. The contribution of our work is two-fold:

1. Our work provides a new perspective to address zero-shot recognition, in which an image feature is disentangled into three latent features via hierarchical structure of disentanglement.

2. With extensive evaluation on popular benchmark datasets, we confirm a significant improvement over
other state-of-the-art methods in both typical and generalized zero-shot recognition.

2. Related work

Our work is related to zero-shot learning and disentangled representation learning. The most widely used intermediate semantic representations in zero-shot learning are called attributes. Human-made attributes [27] are in the form of a category-attribute matrix, each element of which shows if a class has an attribute. Relative attributes [33] capture the semantic relationship, which measures the relative strength of each attribute for unseen classes. Data-driven attributes [47] are a discriminative representation automatically discovered from visual images but are non-interpretable. Word embedding [31] has recently been introduced as a drop-in replacement for attributes [15], which can currently be efficiently trained on a large text corpus.

There are two lines of models for zero-shot recognition based on such attributes. The first line is to learn attribute classifiers for each attribute [27, 22]. Several studies [22, 13] claimed that not all image features are useful for learning a classifier for a given attribute. This problem occurs when a few of the attributes are highly correlated. A feature selection method, such as lasso, is introduced to learn a robust attribute classifier. The second line is to learn a mapping between image features and attributes, and a nearest neighbor search in the mapped space is carried out to predict the class label. There are three different approaches for learning the embedding function: (1) mapping visual features onto the space of intermediate representations [15], (2) mapping the intermediate representations onto the space of visual features [38], and (3) mapping both the visual features and intermediate representations into a common latent space [46, 28, 23]. Most studies learn the embedding function directly from a whole image feature. We do not learn the visual-semantic embedding directly from the whole image feature, since it may hurt the discrimination between seen and unseen classes. In other studies, learning the embedding function and inference are jointly learned in a unified framework [7, 2, 37]. Recent work [44] uses semantic attributes as a condition to generate image features for unseen classes. A classifier is then trained on the generated image features in a supervised manner.

There have been increasing efforts in the deep learning community towards learning factor variations in data in supervised [34, 36], semi-supervised [39, 29] or unsupervised [26] manners. Generative models, such as Boltzmann machine [11], auto-encoder [25] and its variants [20, 30, 24, 42], are widely used for learning disentangled representation from data. From a generative perspective, data is generated via multiplicative interactions of independent factors embedded in the data. InfoGAN [8] introduces disentanglement to a subset of latent variables by maximizing the mutual information between it and the data. Variational autoencoder (VAE) [25, 35] and its variant β-VAE [20] introduce a regularizer of Kullback-Leibler (KL) divergence to a reconstruction error, which pushes the output of the encoder toward a factorial Gaussian prior. Quite recently, total correlation, which measures joint independence for multivariate variables, has been widely used to disentangle features [1, 24, 42]. However, estimations in a mini-batch, such as the Monte-Carlo estimation [24] and the density-ratio estimation [42], are inevitable. These estimations may result in unstable training. Other metrics such as covariance [10] are also used for learning disentangled representation. In addition, most studies that have dimension-wise disentanglement use simple datasets such as celebA and chairs. Independent factors such as azimuth, lighting and elevation can be factorized out. However, the dimension-wise disentanglement for complex real-world images has not been studied due to the fact that independent factors embedded in images might be too difficult to be factorized out with state-of-the-art generative models.

3. The proposal

Image features, such as Res-net features [19], are semantically low-level representations for image concepts. One dimension of such features is strongly correlated with other dimensions due to the spatial interactions across patches from which the convolution operation is performed.

As shown in Figure 2(a), an image feature space can be roughly decomposed into three latent feature spaces from perspectives of discrimination and interpretability, which are semantic, non-semantic and non-discriminative. An overview of our model is shown in Figure 2(b). Our model uses an auto-encoder architecture and factorizes an image feature into these three latent features, which are used for reconstructing the image feature. The semantic latent feature captures discriminative semantic information which is related to human-made attributes. The non-semantic latent feature captures discriminative information that is non-interpretable. The non-discriminative latent feature may consist of intra-class variations and common factors across classes. It tends to have a lower variance of features than the other two.

We denote seen images as $S = \{(x_i, y_i)\}_{i=1}^N$, where $x_i \in \mathbb{R}^d$ is an image feature, $y_i$ is its class label in $\mathcal{Y}^s = \{s_1, s_2, \ldots, s_C\}$ consisting of $C$ class labels, and $N$ is the number of seen images. Each class $s_i$ has a corresponding attribute $a_i \in \mathbb{R}^{d_a}$ in an attribute set $\mathcal{A}^s = \{a_1, a_2, \ldots, a_C\}$. For zero-shot learning, an attribute $a_i$ is typically a per-class attribute vector with continuous values. In the auto-encoder architecture, the encoder learns a mapping from an image feature to a latent factor $z \in \mathbb{R}^l$, from which the decoder reconstructs the image feature. In our case, the latent factor can be a concatenation of the three
latent features, which are the semantic feature $z_s \in \mathbb{R}^{l_s}$, the non-semantic feature $z_n \in \mathbb{R}^{l_n}$, and the non-discriminative feature $z_u \in \mathbb{R}^{l_u}$, such that $z = [z_s, z_n, z_u]$ and $l = l_s + l_n + l_u$. Three main components, which include feature selection, learning non-discriminative features, and hierarchical disentanglement, are used for learning these three latent features. For simplicity, we use $y$ and $a$ in the following sections to represent one-hot vector of class label and class-level attribute for the image feature $x$, respectively.

### 3.1. Feature selection

As suggested by Farhadi et al. [13], not all dimensions of image features are suitable for predicting a specific attribute. It is also intuitive that each latent feature is transformed from disjoint subsets of feature dimensions. For example, dimensions of a feature with low-variance can be used to learn a non-discriminative latent feature, while dimensions with high-variance can be used to learn a discriminative one. Therefore, we learn three transformation matrices that extract different parts of information from an image feature.

We denote transformation matrices $W_s \in \mathbb{R}^{d \times l_s}$, $W_n \in \mathbb{R}^{d \times l_n}$, and $W_u \in \mathbb{R}^{d \times l_u}$ for the semantic feature $z_s$, the non-semantic feature $z_n$, and the non-discriminative feature $z_u$, respectively. We denote $W = [W_s, W_n, W_u]$ such that $W \in \mathbb{R}^{d \times l}$. The objective function for the transformation matrix $W$ is minimized, which is defined as

$$L_{\text{sparse}} = |W|_1 + \lambda \sum_{i \neq j} |w_i^T w_j|$$  \hspace{1cm} (1)$$

where the first term encourages $\text{vec}(W)$ to be sparse, the second term encourages different supports (non-zero elements) for vector pairs, and $\lambda$ is the regularization parameter. The sparse constraint in the first term serves as a feature selector that decides which dimension in an image feature contributes either of three latent features. We squeeze $W_s$, $W_n$, and $W_u$ into vectors $w_1$, $w_2$, and $w_3$, respectively, by having an $l_2$ norm for each row of the transformation matrix. Take the $i$-th element of $w_1$ as an example. It is $l_2$ norm of the $i$-th row of $W_s$. The second term encourages that non-zero elements in any two $w_i(i = 1, 2, 3)$ are different in dimensions. A extreme case is that if non-zero elements in any two $w_i(i = 1, 2, 3)$ are disjoint, the second term becomes zero. Note that Group lasso [22] might not be suitable for our case. This is because group lasso may easily have one group much sparser than other groups, which leads to inappropriate reconstruction for the image feature.

### 3.2. Learning non-discriminative features

Discriminative feature, including semantic and non-semantic, can be extracted by setting a discriminative loss function. However, without learning non-discriminative feature, it does not guarantee that all discriminative information is captured by semantic and non-semantic latent features. Learning the non-discriminative feature is to extract only non-discriminative information from an image feature. The extraction in two directions help decompose discriminative and non-discriminative information.

We learn non-discriminative features using the concept of adversarial learning. Adversarial learning consists of a generator and a discriminator, which are denoted as $G(x) = W_s x$ and $D(z_u) = \sigma(W_u z_u)$, respectively, where $W_c \in \mathbb{R}^{C \times l_u}$, $\sigma$ is the softmax function that outputs the probabilities of class labels, and $z_u$ denotes the output of the generator. The discriminator tries to classify $z_u$ correctly. The generator, however, tries to fool this classifier, which means non-discriminative information is generated from $x$ via $W_u$. The latent feature $z_u$ is likely to have less discriminative information in an image feature. The diagram of adversarial learning is illustrated in Figure 2(c). To be more specific, the value function $V(D, G)$ is as follows:

$$\min_D \max_G \mathbb{E}_{x \sim p(x)} L_c(D(G(x)), y)$$  \hspace{1cm} (2)$$

where $L_c$ is a cross-entropy loss for classification and $y$ is the label of image feature $x$. Learning non-discriminative features can be done by minimizing two functions $L_{\text{adversarial}}^D = L_c(D(G(x)), y)$ and $L_{\text{adversarial}}^G = -L_c(D(G(x)), y)$.
3.3. Hierarchical disentanglement

For a real world image, a dimension of the image feature can not express the most fine-grained factors (concepts) [24, 42]. A group-wise disentanglement is an alternative to capture those fine-grained factors. It is also intuitive that a number of fine-grained factors (low-level concepts) compose a high-level concept, such as tail and head of an animal. The high-level concept helps the generalization to unseen classes. Therefore, besides a group-wise disentanglement, we encourage a hierarchical structure of disentanglement over groups for semantic and non-semantic latent features, as illustrated in Figure 2(d). Note this hierarchical structure is different from [21, 12].

In our model, we propose to use distance covariance [40, 41] to measure the mutual independence between two multivariate variables. Distance covariance is equal to zero if and only if the two variables are mutually independent. This good property is not held by other metrics such as total correlation and covariance. Suppose semantic and non-semantic latent features have the same hierarchical structure. To have concise formulas, here we abuse notation to represent either semantic latent feature or non-semantic latent feature.

In the l-th layer of the hierarchical structure, we have \( z = \{ z_i \}_{i=1}^{c_l} \), where \( z_i \) represents a group of dimensions, and \( c_l \) represents the number of groups in the l-th layer. At each layer, the distance covariances of any different group pairs are averaged. We have a constraint on \( c_l \) such that \( c_l \geq 2 \). The disentanglement loss for the latent feature \( z \) can be written as

\[
L_{\text{disentangle}}^l = \sum_{i} \frac{2}{c_l(c_l - 1)} \sum_{i \neq j} \text{dCov}^2(z_i, z_j), 
\]

(3)

where \( \text{dCov}^2(\cdot, \cdot) \) denotes the distance covariance. Distance covariance \( \text{dCov}^2(h, l) \) can be calculated as follows. Suppose \( H \in \mathbb{R}^{n \times d_h} \) and \( L \in \mathbb{R}^{n \times d_l} \) are two groups of dimensions, where \( d_h \) and \( d_l \) are dimensions of these two groups and \( n \) is the number of observations. Note that \( n \) is the batch size in the experiment. We denote \( a_{ij} = |h_i - h_j| \), where \( i, j = 1, 2, \ldots, n \), and \( h_i, h_j \in \mathbb{R}^{d_h} \). \( \| \cdot \| \) denotes the Euclidean norm, \( \hat{a}_i = \frac{1}{n} \sum_{k=1}^{n} a_{ik} \), \( \hat{a}_j = \frac{1}{n} \sum_{l=1}^{n} a_{lj} \), and \( \hat{a} = \frac{1}{n^2} \sum_{i,l=1}^{n} a_{il} \). Then we have \( A_{i,j} = a_{ij} - \hat{a}_i - \hat{a}_j + \hat{a} \). This also applies to \( L \), and we then have \( B_{i,j} = b_{ij} - \hat{b}_i - \hat{b}_j + \hat{b} \). Finally, we have \( \text{dCov}^2(h, l) = \frac{1}{n^2} \sum_{i,l=1}^{n} A_{i,j} B_{i,j} \). An example is given as illustrated in Figure 2(d). Suppose \( z \) is composed of four groups \( \{ g_1, g_2, g_3, g_4 \} \), we can have the loss for \( z \) by summing the losses of the first and second layers, which are denoted by \( L_{d_1} \) and \( L_{d_2} \), respectively. There is no loss for the third layer because only one group exists. We can easily have \( L_{d_1} = \frac{1}{2} \sum \text{dCov}^2(g_i, g_j) \) \((i \neq j)\) and \( L_{d_2} = \text{dCov}^2(\{ g_1, g_2 \}, \{ g_3, g_4 \}) \).

We have a loss function for both semantic and non-semantic latent features, which is represented by \( L_{\text{disentangle}} = L_{\text{disentangle}}^s + L_{\text{disentangle}}^n \). \( L_{\text{disentangle}}^s \) and \( L_{\text{disentangle}}^n \) denote the losses calculated by Equation 3 for semantic and non-semantic latent features, respectively.

3.4. Learning and inference

We introduce the other components for disentangling the image feature. They are (1) reconstruction loss \( L_{\text{recon}} \); this loss is minimized, which is denoted as \( L_{\text{recon}} = \| x - Vz \| \), where \( V \in \mathbb{R}^{d \times l} \) is a transformation matrix and \( z = [z_s, z_n, z_a] \); (2) classification loss \( L_{\text{classification}} \); both semantic and non-semantic latent features are responsible for classification. Cross entropy loss is calculated from \( \sigma(W_dz_d) \) and ground-truth class label, where \( \sigma \) is the softmax function, \( W_d \in \mathbb{R}^{l \times (l + l_n)} \), and \( z_d = [z_s, z_n] \); (3) regularization loss \( L_{\text{regularization}} \); this loss function encourages the semantic latent feature semantically close to human-made attributes. This is achieved by minimizing \( L_{\text{regularization}} = \| a - W_a z_s \| \) where \( W_a \in \mathbb{R}^{d_a \times l_s} \) is a transformation matrix.

By using the auto-encoder framework shown in Figure 2(b), the image feature is disentangled into three parts, which are semantic \( (z_s) \), non-semantic \( (z_n) \) and non-discriminative \( (z_a) \). The fusion of \( z_s \) and \( z_n \), which is taken a variant of image feature, is further used for learning the visual-semantic embedding function. The fusion could be a simple concatenation or a more sophisticated compact bilinear pool [17]. For learning the embedding function, we can use any state-of-the-arts such as DeViSE [15]. We denote the objective function of learning the embedding function as \( L_{\text{embedding}} \), which is a margin hinge loss in DeViSE. The fundamental idea in DeViSE is that the similarity between a latent feature of a class and its attribute representation should be larger than that between a latent feature of another class and the attribute representation. Learning three latent features and embedding function can be trained jointly. In the inference stage, semantic and non-semantic latent features are extracted from an unseen image, and these two latent features are mapped to a visual-semantic space. The unseen image is assigned to the label of the closest attribute representation via nearest-neighbor search in that space.

We observed through experiments that simply combining all loss functions for disentangling feature often results in poor performance. Therefore, we optimize the parameters of our model by jointly learning the following loss functions, i.e., \( L, L_{\text{disentangle}}, L_{\text{adversarial}}^s, L_{\text{adversarial}}^n \) and \( L_{\text{embedding}} \). \( L = L_{\text{recon}} + \alpha L_{\text{sparse}} + \beta L_{\text{classification}} + \gamma L_{\text{regularization}} \), where \( \alpha, \beta, \) and \( \gamma \) are regularization parameters. Our model is not sensitive to parameter settings. In our experiments, superior performances were achieved by simply setting the same values of \( \alpha, \beta, \) and \( \gamma \) for all data sets.
4. Experiments

4.1. Experimental Settings

We used the most widely used data sets for evaluating the performance of zero-shot learning. The datasets were Animals with Attribute (AwA), CUB-200-2011 (CUB), SUN with Attribute (SUN) and Attribute Pascal and Yahoo (aPY). We used new partitions for the above data sets proposed by Xian et al. [45] for a fair comparison, since some classes of images are unfairly used for training a CNN model. We used Res-net-101 [19] as the image features for both seen and unseen classes. Following [45], we used the average top-1 recognition accuracy of each class to evaluate the performance of typical zero-shot learning. For generalized zero-shot learning, we used the harmonic mean, which is defined as $2 \times (c_u \times c_c)/(c_u + c_c)$, where $c_u$ and $c_c$ are the top-1 recognition accuracies for seen and unseen images, respectively.

Since the embedding function in DeVise [15] is a simple yet effective embedding model, we used it as the baseline embedding function to compete with other sophisticated methods for zero-shot learning. We call our model Discriminative Latent Features for Zero-shot Learning (DLFZRL), which includes latent-feature learning and embedding function in DeVise. DLFZRL+w/o D denotes our full model without using disentanglement. DLFZRL+w/o HD denotes our model using group-wise disentanglement but without a hierarchical structure of disentanglement. The setting of the parameters shared by all datasets was as follows. Adam was used as the optimizer, and its initial learning rate was set to $2 \times 10^{-4}$. The activation function after the transformations from the image feature to $z_s, z_n$, and $z_u$ was set to ReLU. The batch size was set to 128. The $\lambda$ value was set to $2 \times 10^{-5}$. The $\alpha, \beta$, and $\gamma$ values were set to 0.01, 1.0, and 1.0, respectively. The dimensions of features for $z_s, z_n$, and $z_u$ were set to 512. The hierarchical structure of disentanglement has two layers, in which the dimension of a group in the first and second layers was set to 16 and 64, respectively. For these partition, we made a grid search for partitions in sizes of exponents of two and chose the best ones. For learning the embedding in DeVise, the margin was set to 0.1.

4.2. Results on image recognition

Table 1 shows the performances of our model and its variants in the benchmark data sets. f-CLSWGAN [44] is the model including a variant of an improved WGAN (f-WGAN) [18] and a softmax classifier. Seen image features and class-level attributes are used to train f-WGAN and image features of unseen classes can be generated given unseen class-level attributes. The softmax classifier with a one-layer neural network is trained for classifying the generated unseen images in a supervised manner. DLFZRL+softmax denotes the model in which we train f-WGAN by using semantic and non-semantic latent features and class-level attributes from seen classes, and the softmax classifier as [44] is used. For fair comparison, we directly cite the results in [45][44], except DLFZRL and its variants. Due to limited space, we only cited methods that tend to achieve best performances in [45][44]. Among all methods, DLFZRL tended to perform the best on most data sets, no matter if the setting was traditional or generalized zero-shot learning. Compared with traditional zero-shot learning, the improvements achieved for generalized zero-shot learning are more significant. The performance improved by 18.1, 17.3 and 21.8 points for AwA1, AwA2 and aPY, respectively. Note that f-CLSWGAN is a method for converting the problem of zero-shot learning into a supervised-learning problem, making it a transductive method that can access the manifold of unseen classes. Direct comparison of its performance with embedding methods is not fair, as they are inductive and do not use the manifold of unseen classes. To verify the effectiveness of semantic and non-semantic latent features, we train f-WGAN by using our latent features and generate features of unseen classes given unseen class-level attributes. We can see that DLFZRL+softmax was superior to f-CLSWGAN in all cases. We also observed that, without group-wise disentanglement or hierarchical disentanglement over groups, DLFZRL+w/o D and DLFZRL+w/o HD were inferior to DLFZRL for both tra-
4.3. Analysis and Discussion

We examine if the non-discriminative latent feature captures low-variance information in the image feature. Figure 3(a) shows the variance of each dimension in descending order for each latent feature in the CUB data set. Figure 3(b) shows the average variance for each latent feature in the CUB data set. Figure 4.3 shows the average variance for each latent feature in the CUB data set. Figure 6.3 shows the degree of discrimination between seen and unseen classes. As shown in Figure 1(a), features of seen classes are less discriminative to that of unseen classes in the feature space. We calculate the ratio of between-class scatter to within-class scatter for image features and the combination of semantic and non-semantic latent features. The larger the ratio, the more seen classes are discriminative from unseen classes. The group-wise disentanglement and hierarchical disentanglement over groups for both semantic and non-semantic latent features improved the degree of discrimination compared with using whole image features.

Table 2 shows the impact of three components, which are 1) feature selection, 2) learning non-discriminative features, and 3) hierarchical disentanglement. We can see that each component does contribute to the performance improvement.
Table 2. Ablation study on three main components in DLFZRL. Bold font denotes the most significant drop in performance. w/o FL denotes DLFZRL without using feature selection, i.e., Equation 1; w/o ND denotes DLFZRL without learning non-discriminative features, i.e., Equation 2; w/o HD denotes DLFZRL without learning hierarchical disentanglement, i.e., Equation 3.

<table>
<thead>
<tr>
<th>Method</th>
<th>SUN</th>
<th>CUB</th>
<th>AWA1</th>
<th>AWA2</th>
<th>aPY</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLFZRL</td>
<td>59.3</td>
<td>57.8</td>
<td>66.3</td>
<td>63.7</td>
<td>44.5</td>
</tr>
<tr>
<td>w/o FL</td>
<td>58.3</td>
<td>55.6</td>
<td>62.9</td>
<td><strong>62.5</strong></td>
<td>42.2</td>
</tr>
<tr>
<td>w/o ND</td>
<td>58.6</td>
<td>56.1</td>
<td>64.3</td>
<td>63.3</td>
<td><strong>41.5</strong></td>
</tr>
<tr>
<td>w/o HD</td>
<td><strong>57.2</strong></td>
<td><strong>55.1</strong></td>
<td><strong>57.9</strong></td>
<td>63.1</td>
<td>41.6</td>
</tr>
</tbody>
</table>

**zero-shot learning (ZRL)**

**generalized zero-shot learning (GZRL)**

<table>
<thead>
<tr>
<th>Method</th>
<th>SUN</th>
<th>CUB</th>
<th>AWA1</th>
<th>AWA2</th>
<th>aPY</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLFZRL</td>
<td>24.6</td>
<td>37.1</td>
<td>40.5</td>
<td>45.1</td>
<td>31.0</td>
</tr>
<tr>
<td>w/o FL</td>
<td><strong>23.2</strong></td>
<td><strong>33.1</strong></td>
<td><strong>36.3</strong></td>
<td>41.2</td>
<td><strong>27.1</strong></td>
</tr>
<tr>
<td>w/o ND</td>
<td>23.8</td>
<td>34.3</td>
<td>38.4</td>
<td>40.9</td>
<td>28.5</td>
</tr>
<tr>
<td>w/o HD</td>
<td><strong>23.2</strong></td>
<td>35.4</td>
<td>37.1</td>
<td><strong>40.6</strong></td>
<td>30.7</td>
</tr>
</tbody>
</table>

Figure 8. Mutual information between two groups in the two layers. Both x-axis and y-axis denote the number of groups. The number of groups in Figure 8(a) and Figure 8(b) are 32 and 8, respectively.

Figure 7. Relation between aPY’s attributes and two layers of groups in semantic latent feature. Figure 7(a) and Figure 7(b) denote the relations in the first and second layers, respectively. The x-axis and y-axis denote the number of groups and attributes, respectively. The number of attributes in aPY is 64. The number of groups in Figure 7(a) and Figure 7(b) are 32 and 8, respectively.

In ZRL, the impact of components is ordered as HD > FL > ND, while the order is FL > HD > ND in GZRL. We can see that, in zero-shot learning, transferable feature learned by HD from seen classes is essential for classifying unseen classes. However, in GZRL, feature selection, which can separate seen and unseen classes, is more important than in ZRL. The average sparse ratios in FL, which is the ratio of the number of near-zero entries to the number of entries in W, over different data sets in ZRL and GZRL are 0.49 and 0.54, respectively. This suggests that FL plays more important role in GZRL and ZRL.

For a 512-dimensional semantic latent feature, suppose the hierarchical structure of disentanglement has two layers. The first layer has 32 groups, each of which has 16 dimension (16D group). The second layer has 8 groups, each of which has 64 dimension (64D group).

First, we examine if the groups in both two layers are able to capture information in attributes. Figure 7 shows the relation between attributes and groups in the two layers for the aPY data set. Mutual information (MI) measures how the groups in z_s are related to the attributes. The higher the value, the more a group is related to an attribute. For a better visualization, we have blocks black when MI is smaller than a threshold. We can observe that some groups relate to disjoint subsets of attributes, while others share attributes. This implies that groups in both two layers are able to capture information of specific attributes.

Second, we examine if the hierarchical structure of groups is appropriately learned in the two layers. Figure 8 illustrates MI between any two groups in each layer for the semantic latent feature in the aPY data set. Group-wise disentanglement in each layer encourages the information of two different groups in z_s independent, while that in the same group dependent. Independent groups correspond to black blocks that have low values of MI, while dependent groups correspond to white blocks that have high values of MI. The blocks along the main diagonal in Figure 8(a) and 8(b) verified that the hierarchical disentanglement is appropriately learned.

5. Conclusion

In zero-shot recognition, most existing methods have a limited capability of recognizing unseen classes. This incapability is due to that both image features trained with seen classes and man-made attributes are variant to unseen classes. In this paper, we propose a model that factorizes an image feature into three latent features, which are called semantic, non-semantic, and non-discriminative. A group-wise disentanglement and hierarchical structure of disentanglement over groups are encouraged for both semantic and non-semantic features. In extensive experiments, superior performances confirmed the effectiveness of the latent features for traditional and generalized zero-shot learning.
References


