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

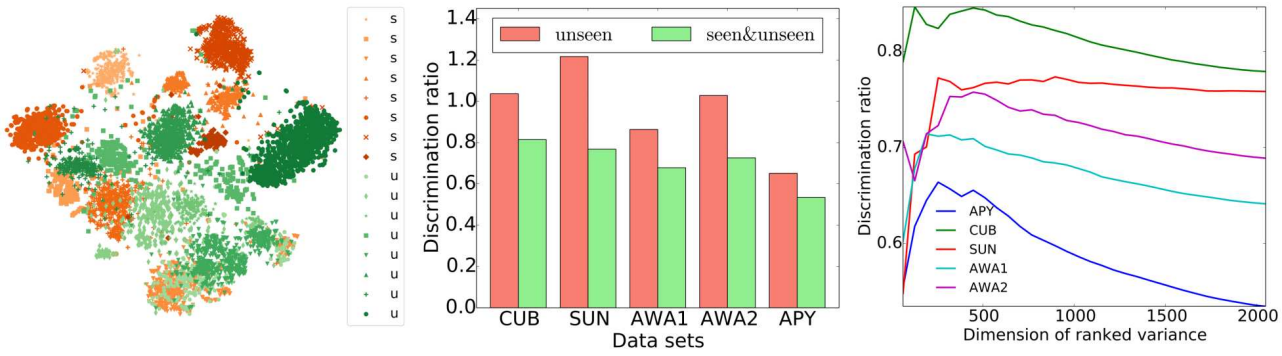
Most studies in zero-shot learning model the relationship, in the form of a classifier or mapping, between features from images of seen classes and their attributes. Therefore, the degree of a model's generalization ability for recognizing unseen images is highly constrained by that of image features 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.



(a) (b) (c)

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 *seen* and *unseen* classes. Discrimination between *seen* and *unseen* classes is measured by the ratio of between-class scatter (BS) to within-class scatter. The bigger the ratio is, the more discriminative the seen and unseen classes are. Figure 1(b) is a histogram of discrimination ratio for (1) unseen classes and (2) both seen & unseen classes. In case (1), BS is the average of the distances between centroids of each unseen class and its nearest class. In case (2), BS is the average of the distances between centroids of each unseen class and its nearest seen class, which was also used for plotting Figure 1(c). In Figure 1(c), the ranked variance is calculated as follows. Suppose the data matrix $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \in \mathbb{R}^{N \times d}$, where N is the number of data and d denotes the dimension of features. The variance σ_i is calculated for the i-th column, i.e., $\mathbf{x}_i$. The variances are ranked in descending order and columns $\mathbf{x}_i$ 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 $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^d$ 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 $\mathbf{a}_i \in \mathbb{R}^{d_a}$ in an attribute set $\mathcal{A}^s = \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_C\}$. For zero-shot learning, an attribute $\mathbf{a}_i$ 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 $\mathbf{z} \in \mathbb{R}^l$, from which the decoder reconstructs the image feature. In our case, the latent factor can be a concatenation of the three

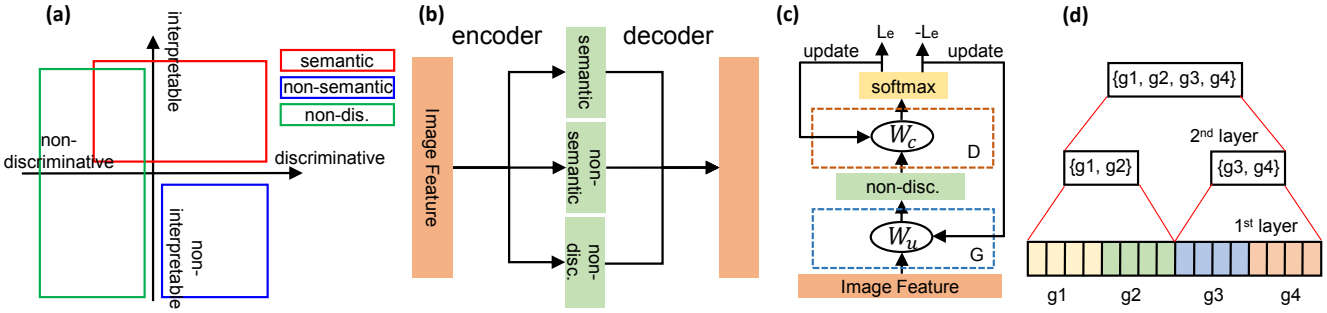


Figure 2. (a) Three latent features in the space with interpretability and discrimination axes. (b) Framework of our model that factorizes an image feature into three latent features. (c) Adversarial learning for a non-discriminative latent feature. (d) Hierarchical structure for a latent feature.

latent features, which are the *semantic* feature $\mathbf{z}_s \in \mathbb{R}^{l_s}$, the *non-semantic* feature $\mathbf{z}_n \in \mathbb{R}^{l_n}$, and the *non-discriminative* feature $\mathbf{z}_u \in \mathbb{R}^{l_u}$, such that $\mathbf{z} = [\mathbf{z}_s, \mathbf{z}_n, \mathbf{z}_u]$ 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 $\mathbf{y}$ and $\mathbf{a}$ in the following sections to represent one-hot vector of class label and class-level attribute for the image feature $\mathbf{x}$, 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 $\mathbf{W}_s \in \mathbb{R}^{d \times l_s}$, $\mathbf{W}_n \in \mathbb{R}^{d \times l_n}$ and $\mathbf{W}_u \in \mathbb{R}^{d \times l_u}$ for the *semantic* feature $\mathbf{z}_s$, the *non-semantic* feature $\mathbf{z}_n$, and the *non-discriminative* feature $\mathbf{z}_u$, respectively. We denote $\mathbf{W} = [\mathbf{W}_s, \mathbf{W}_n, \mathbf{W}_u]$ such that $\mathbf{W} \in \mathbb{R}^{d \times l}$. The objective function for the transformation matrix $\mathbf{W}$ is minimized, which is defined as

$$L_{\text{sparse}} = |\mathbf{W}|_1 + \lambda \sum_{i \neq j} |\mathbf{w}_i^T \mathbf{w}_j| \tag{1}$$

where the first term encourages $\mathbf{vec}(\mathbf{W})$ to be sparse, the second term encourages different supports (non-zero elements) for vector pairs, and λ 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 $\mathbf{W}_s$, $\mathbf{W}_n$ and $\mathbf{W}_u$ into vectors $\mathbf{w}_1$, $\mathbf{w}_2$, and $\mathbf{w}_3$, respectively, by having a l_2 norm for each row of the transformation matrix. Take the i-th element of $\mathbf{w}_1$ as an example. It is l_2 norm of the i-th row of $\mathbf{W}_s$. The second term encourages that non-zero elements in any two $\mathbf{w}_i (i = 1, 2, 3)$ are different in dimensions. A extreme case is that if non-zero elements in any two $\mathbf{w}_i (i = 1, 2, 3)$ 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 $G(\mathbf{x}) = \mathbf{W}_u \mathbf{x}$ and $D(\mathbf{z}_u) = \sigma(\mathbf{W}_c \mathbf{z}_u)$, respectively, where $\mathbf{W}_c \in \mathbb{R}^{C \times l_u}$, σ is the softmax function that outputs the probabilities of class labels, and $\mathbf{z}_u$ denotes the output of the generator. The discriminator tries to classify $\mathbf{z}_u$ correctly. The generator, however, tries to fool this classifier, which means non-discriminative information is generated from $\mathbf{x}$ via $\mathbf{W}_u$. The latent feature $\mathbf{z}_u$ 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:

$$\min_D \max_G V(D, G) = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})} L_e(D(G(\mathbf{x})), \mathbf{y}) \tag{2}$$

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

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 $\mathbf{z}$ to represent either *semantic* latent feature or *non-semantic* latent feature.

In the l-th layer of the hierarchical structure, we have $\mathbf{z} = \{\mathbf{z}_l^i\}_{i=1}^{c_l}$, where $\mathbf{z}_l^i$ 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 $\mathbf{z}$ can be written as

$$L'_{\text{disentangle}} = \sum_l \frac{2}{c_l(c_l-1)} \sum_{i \neq j}^{c_l} \text{dCov}^2(\mathbf{z}_l^i, \mathbf{z}_l^j), \tag{3}$$

where $\text{dCov}^2(\cdot,\cdot)$ denotes the distance covariance. Distance covariance $\text{dCov}^2(\mathbf{h}, \mathbf{l})$ can be calculated as follows. Suppose $\mathbf{H} \in \mathbb{R}^{n \times d_h}$ and $\mathbf{L} \in \mathbb{R}^{n \times d_l}$ 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 $a_{ij} = |\mathbf{h}_i - \mathbf{h}_j|$, where $i, j = 1, 2, \ldots, n$, $\mathbf{h}_i, \mathbf{h}_j \in \mathbb{R}^{d_h}$, $|\cdot|$ 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 $\hat{a} = \frac{1}{n^2}\sum_{k,l=1}^n a_{kl}$. Then we have $\mathbf{A}_{i,j} = a_{ij} - \hat{a}_i - \hat{a}_j + \hat{a}$. This also applies to $\mathbf{L}$, and we then have $\mathbf{B}_{i,j} = b_{ij} - \hat{b}_i - \hat{b}_j + \hat{b}$. Finally, we have $\text{dCov}^2(\mathbf{h}, \mathbf{l}) = \frac{1}{n^2}\sum_{i,j}^n \mathbf{A}_{i,j}\mathbf{B}_{i,j}$. An example is given as illustrated in Figure 2(d). Suppose $\mathbf{z}$ is composed of four groups $\{g_1, g_2, g_3, g_4\}$, we can have the loss for $\mathbf{z}$ 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 $L_{d_1} = \frac{1}{6}\sum \text{dCov}^2(g_i, g_j)$ (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 $L_{\text{disentangle}} = L^s_{\text{disentangle}} + L^n_{\text{disentangle}}$. $L^s_{\text{disentangle}}$ and $L^n_{\text{disentangle}}$ 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_{recon}: this loss is minimized, which is denoted as $L_{\text{recon}} = ||\mathbf{x} - \mathbf{V}\mathbf{z}||$, where $\mathbf{V} \in \mathbb{R}^{d \times l}$ is a transformation matrix and $\mathbf{z} = [\mathbf{z}_s, \mathbf{z}_n, \mathbf{z}_u]$; (2) classification loss $L_{\text{classification}}$: both *semantic* and *non-semantic* latent features are responsible for classification. Cross entropy loss is calculated from $\sigma(\mathbf{W}_d\mathbf{z}_d)$ and ground-truth class label, where σ is the softmax function, $\mathbf{W}_d \in \mathbb{R}^{C \times (l_s + l_n)}$, and $\mathbf{z}_d = [\mathbf{z}_s, \mathbf{z}_n]$; (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 $L_{\text{regularization}} = ||\mathbf{a} - \mathbf{W}_a\mathbf{z}_s||$ where $\mathbf{W}_a \in \mathbb{R}^{d_a \times l_s}$ 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* ($\mathbf{z}_s$), *non-semantic* ($\mathbf{z}_n$) and *non-discriminative* ($\mathbf{z}_u$). The fusion of $\mathbf{z}_s$ and $\mathbf{z}_n$, 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 pooling [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 combing 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, $L_{\text{disentangle}}$, $L^D_{\text{adversarial}}$, $L^G_{\text{adversarial}}$, and $L_{\text{embedding}}$. $L = L_{\text{recon}} + \alpha L_{\text{sparse}} + \beta L_{\text{classification}} + \gamma L_{\text{regularization}}$, where α, β, and γ 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 α, β, and γ for all data sets.

Table 1. Performances of models in popular benchmark data sets.

	traditional zero-shot learning					generalized zero-shot learning				
Method	SUN	CUB	AWA1	AWA2	aPY	SUN	CUB	AWA1	AWA2	aPY
DeViSE[15]	56.5	52.0	54.2	59.7	39.8	20.9	32.8	22.4	27.8	9.2
ALE[3]	58.1	54.9	59.9	62.5	39.7	26.3	34.4	27.5	23.9	8.7
SJE[4]	53.7	53.9	65.6	61.9	32.9	19.8	33.6	19.6	14.4	6.9
DLFZRL+w/o D	55.6	53.4	58.9	61.5	39.1	21.7	34.2	34.8	38.4	23.1
DLFZRL+w/o HD	57.2	55.1	57.9	63.1	41.6	23.2	35.4	37.1	40.6	30.7
DLFZRL	**59.3**	**57.8**	**66.3**	**63.7**	**44.5**	24.6	**37.1**	**40.5**	**45.1**	**31.0**
f-CLSWGAN[44]	60.8	57.3	68.2	-	-	39.4	49.7	59.6	-	-
DLFZRL+softmax	**61.3**	**61.8**	**71.3**	**70.3**	**46.7**	**42.5**	**51.9**	**61.2**	**60.9**	**38.5**

4. Experiments

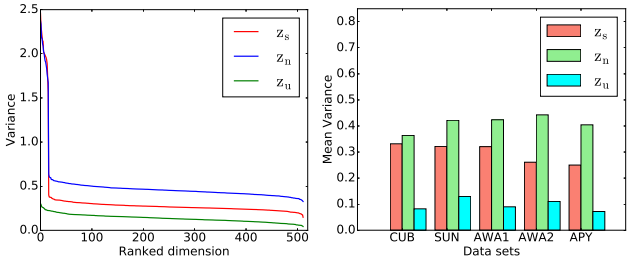
4.1. Experimental Settings

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 $2 \times (c_{\text{tr}} \times c_{\text{te}})/(c_{\text{tr}} + c_{\text{te}})$, where c_{tr} and c_{te} 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 2×10^{-4}. The activation function after the transformations from the image feature to $\mathbf{z}_s$, $\mathbf{z}_n$, and $\mathbf{z}_u$ was set to ReLU. The batch size was set to 128. The λ value was set to 2×10^{-3}. The α, β, and γ values were set to 0.01, 1.0, and 1.0, respectively. The dimensions of features for $\mathbf{z}_s$, $\mathbf{z}_n$, and $\mathbf{z}_d$ 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.

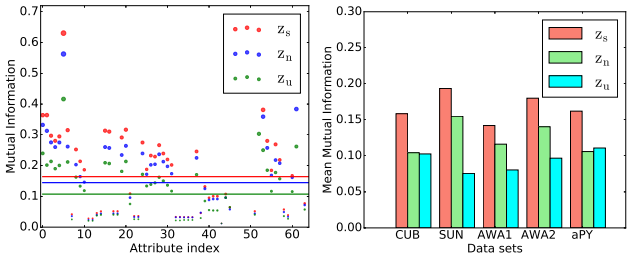
4.2. Results on image recognition

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-



(a) CUB (b) All data sets

Figure 3. Variance in each latent feature. The *non-discriminative* latent feature $\mathbf{z}_u$ has the lowest variance. The *semantic* and *non-semantic* latent features, i.e., $\mathbf{z}_s$ and $\mathbf{z}_n$, have much higher variances than the *non-discriminative* latent feature.



(a) aPY (b) All data sets

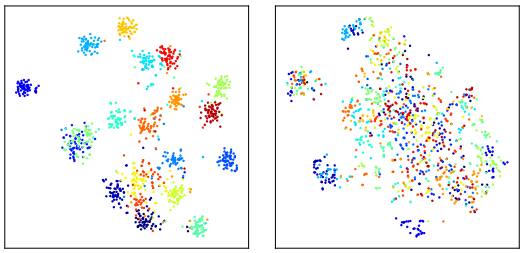
Figure 4. Relatedness of attributes. Relatedness is measured by mutual information between each dimension of latent feature and attribute. In Figure 4(a), the average relatedness of each attribute is represented by a horizontal line for each latent feature.

ditional and generalized zero-shot learning. This implies that group-wise disentanglement and hierarchical disentanglement over groups are helpful for learning discriminative and more generalizable latent features. DLFZRL outperformed current methods in most cases, which further confirms the superiority of DLFZRL.

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 different data sets. We can see that the *non-semantic* latent feature had even higher variance than the *semantic* latent feature. This is consistent with the intuition that the human-made attributes are discriminative to some extent, while the *non-semantic* latent feature is data-driven, which tends to be more discriminative.

To examine if $\mathbf{z}_s$ is able to encode the information in attribute, we show in Figure 4(a) the relatedness of each attribute with respect to each latent feature in the aPY data set. Mutual information (MI) between two vectors is calculated by $I(\mathbf{x},\mathbf{y}) = \sum_{x\epsilon\mathbf{x}}\sum_{y\epsilon\mathbf{y}} p(x,y)log\frac{p(x,y)}{p(x)p(y)}$, where



(a) $\mathbf{z}_n$ (b) $\mathbf{z}_u$

Figure 5. t-SNE visualization of $\mathbf{z}_n$ and $\mathbf{z}_u$. Different colors denote different classes.

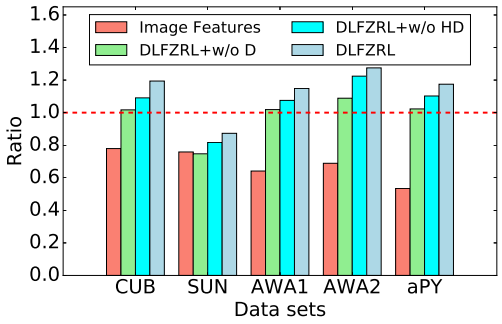


Figure 6. Discrimination ratios of different methods in different data sets.

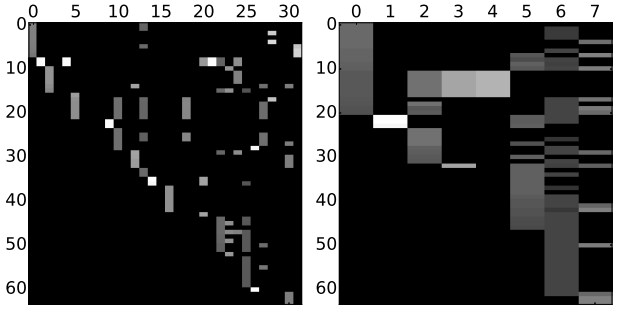
x and y are values in the vectors $\mathbf{x}$ and $\mathbf{y}$, respectively. This figure verifies that the *semantic* latent feature is the most related to attributes compared with the other two latent features, showing the effectiveness of regularization on attributes. Figure 4(b) shows the relatedness of attributes in different data sets. The *semantic* latent feature always achieves the highest relatedness.

Figure 5 illustrates the latent features $\mathbf{z}_n$ and $\mathbf{z}_u$ for *unseen* classes in the CUB data set by using t-SNE [43]. *Unseen* classes are discriminative in $\mathbf{z}_n$ as shown in Figure 5(a), while they have no discrimination in $\mathbf{z}_u$ as shown in Figure 5(b). Figure 6 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 im-

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.

zero-shot learning (ZRL)					
Method	SUN	CUB	AWA1	AWA2	aPY
DLFZRL	59.3	57.8	66.3	63.7	44.5
w/o FL	58.3	55.6	62.9	**62.5**	42.2
w/o ND	58.6	56.1	64.3	63.3	**41.5**
w/o HD	**57.2**	**55.1**	**57.9**	63.1	41.6
generalized zero-shot learning (GZRL)					
DLFZRL	24.6	37.1	40.5	45.1	31.0
w/o FL	**23.2**	**33.1**	**36.3**	41.2	**27.1**
w/o ND	23.8	34.3	38.4	40.9	28.5
w/o HD	**23.2**	35.4	37.1	**40.6**	30.7



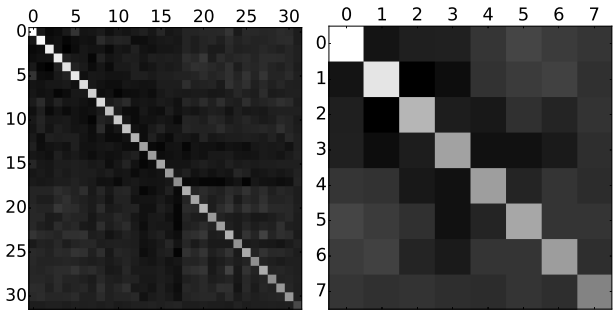
(a) 16D group (b) 64D group

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.

provement. 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 $\mathbf{W}$, 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



(a) 16D group (b) 64D group

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.

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 $\mathbf{z}_s$ 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 $\mathbf{z}_s$ 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

[1] Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE TPAMI*, (99):1–1, 2016.

[2] Zeynep Akata, Florent Perronnin, Zaïd Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *CVPR*, pages 819–826, 2013.

[3] Zeynep Akata, Florent Perronnin, Zaïd Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE TPAMI*, 38(7):1425–1438, 2016.

[4] Zeynep Akata, Scott E. Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, pages 2927–2936, 2015.

[5] Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE TPAMI*, 35(8):1798–1828, 2013.

[6] Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors. *Zero-Shot Learning Through Cross-Modal Transfer*, 2013.

[7] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, pages 5327–5336, 2016.

[8] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, 2016.

[9] Christopher K. I. Williams Cian Eastwood. A framework for the quantitative evaluation of disentangled representations. In *ICLR*, 2018.

[10] Michael Cogswell, Faruk Ahmed, Ross B. Girshick, Larry Zitnick, and Dhruv Batra. Reducing overfitting in deep networks by decorrelating representations. In *ICLR*, 2016.

[11] Guillaume Desjardins and Aaron Courville amd Yoshua Bengio. Disentangling factors of variation via generative entangling. *arXiv preprint arXiv:1210.5474*, 2012.

[12] Babak Esmaeili, Hao Wu, Sarthak Jain, Dana H. Brooks, Jennifer Dy, and Jan-Willem van de Meent. Structured disentangled representations. *arXiv:1804.02086*, 2018.

[13] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *CVPR*, pages 1778–1785, 2009.

[14] Rafael Felix, B. G. Vijay Kumar, Ian D. Reid, and Gustavo Carneiro. Multi-modal cycle-consistent generalized zero-shot learning. In *ECCV*, pages 21–37, 2018.

[15] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129, 2013.

[16] Yanwei Fu, Tao Xiang, Yu-Gang Jiang, Xiangyang Xue, Leonid Sigal, and Shaogang Gong. Recent advances in zero-shot recognition. *arXiv preprint arXiv:1710.04837*, 2017.

[17] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *CVPR*, pages 317–326, 2016.

[18] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. In *NIPS*, pages 5769–5779, 2017.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[20] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

[21] Irina Higgins, Nicolas Sonnerat, Loïc Matthey, Arka Pal, Christopher Burgess, Matthew Botvinick, Demis Hassabis, and Alexander Lerchner. Scan: Learning abstract hierarchical compositional visual concepts. In *ICLR*, 2018.

[22] Dinesh Jayaraman, Fei Sha, and Kristen Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *CVPR*, pages 1629–1636, 2014.

[23] Huajie Jiang, Ruiping Wang, Shiguang Shan, Yi Yang, and Xilin Chen. Learning discriminative latent attributes for zero-shot classification. In *ICCV*, pages 4233–4242, 2017.

[24] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.

[25] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[26] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *ICLR*, 2018.

[27] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, pages 951–958, 2009.

[28] Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors. *ECCV*, 2016.

[29] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian J. Goodfellow. Adversarial autoencoders. In *ICLR*, 2016.

[30] Michaël Mathieu, Junbo Jake Zhao, Pablo Sprechmann, Aditya Ramesh, and Yann LeCun. Disentangling factors of variation in deep representations using adversarial training. In *NIPS*, pages 5040–5048, 2016.

[31] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR Workshop*, 2013.

[32] Norouzi Mohammad, Mikolov Tomas, Bengio Samy, Singer Yoram, Shlens Jonathon, Frome Andrea, Corrado Greg, and Dean Jeffrey. Zero-shot learning by convex combination of semantic embeddings. In *ICLR*, 2014.

[33] Devi Parikh and Kristen Grauman. Relative attributes. In *ICCV*, pages 503–510, 2011.

[34] Scott E. Reed, Kihyuk Sohn, Yuting Zhang, and Honglak Lee. Learning to disentangle factors of variation with manifold interaction. In *ICML*, pages 1431–1439, 2014.

[35] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, pages 1278–1286, 2014.

[36] Karl Ridgeway and Michael C. Mozer. Learning deep disentangled embeddings with the f-statistic loss. *arXiv preprint arXiv:1802.05312*, 2018.

[37] Bernardino Romera-Paredes and Philip H. S. Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, pages 2152–2161, 2015.

[38] Yutaro Shigeto, Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, and Yuji Matsumoto. Ridge regression, hubness, and zero-shot learning. In *ECML-PKDD*, pages 135–151, 2015.

[39] N. Siddharth, Brooks Paige, Jan Willem Van De Meent, Alban Desmaison, Noah D. Goodman, Pushmeet Kohli, Frank Wood, and Philip H. S. Torr. Learning disentangled representations with semi-supervised deep generative models. In *NIPS*, 2017.

[40] GáBor J. SzéKely, Maria Rizzo L., and Nail Bakirov K. Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35, 2007.

[41] GáBor J. SzéKely and Maria L. Rizzo. The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*, 117:193–213, May 2013.

[42] Roger Grosse David Duvenaud Tian Qi Chen, Xuechen Li. Isolating sources of disentanglement in variational autoencoders. *arXiv preprint arXiv:1802.04942*, 2018.

[43] L.J.P van der Maaten and G.E. Hinton. Visualizing high-dimensional data using t-sne. *JMLR*, 9:2579–2605, 2008.

[44] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, 2018.

[45] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning - the good, the bad and the ugly. In *CVPR*, pages 3077–3086, 2017.

[46] Y. Yang and T. M. Hospedales. A unified perspective on multi-domain and multi-task learning. In *ICLR*, 2015.

[47] Felix X. Yu, Liangliang Cao, Rogério Schmidt Feris, John R. Smith, and Shih-Fu Chang. Designing category-level attributes for discriminative visual recognition. In *CVPR*, pages 771–778, 2013.

[48] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. In *ICCV*, pages 4166–4174, 2015.