Recognizing Unseen Attribute-Object Pair with Generative Model

Zhixiong Nan[1,2*] **, Yang Liu**[2*]**, Nanning Zheng**[1]**, and Song-Chun Zhu**[2]
[1]Xi'an Jiaotong University, China
[2]University of California, Los Angeles, USA

Abstract

In this paper, we are studying the problem of recognizing attribute-object pairs that do not appear in the training dataset, which is called unseen attribute-object pair recognition. Existing methods mainly learn a discriminative classifier or compose multiple classifiers to tackle this problem, which exhibit poor performance for unseen pairs. The key reasons for this failure are 1) they have not learned an intrinsic attribute-object representation, and 2) the attribute and object are processed either separately or equally so that the inner relation between the attribute and object has not been explored. To explore the inner relation of attribute and object as well as the intrinsic attribute-object representation, we propose a generative model with the encoder-decoder mechanism that bridges visual and linguistic information in a unified end-to-end network. The encoder-decoder mechanism presents the impressive potential to find an intrinsic attribute-object feature representation. In addition, combining visual and linguistic features in a unified model allows to mine the relation of attribute and object. We conducted extensive experiments to compare our method with several state-of-the-art methods on two challenging datasets. The results show that our method outperforms all other methods.

Introduction

Attributes are the description of 'objects'. To reach a higher level of vision understanding, a computer should understand not only object categories but also their attributes. As a result, recognizing objects with their attributes have been widely studied in various problems such as person re-identification (Su et al. 2016), scene understanding (Laffont et al. 2014), image caption (Wu et al. 2017), image search (Kumar, Belhumeur, and Nayar 2008), and image generation (Yan et al. 2016).

Encouraged by the success of discriminative models implemented with deep neural networks for object classification (Simonyan and Zisserman 2014; He et al. 2016; Huang et al. 2017), some studies like (Misra, Gupta, and Hebert 2017) have tried to recognize attribute-object pairs by composing the discriminative models that are separately trained for the object and attribute. Factually, the discriminative models are trying to learn the attribute visual 'prototype' and object visual 'prototype'. It is true that an object always has a visual 'prototype'. For example, when we ask people to draw a dog, different people may draw beagles, collies, dalmatians or poodles, but the 'dogs' they draw always have two ears and four legs. However, if we ask people to draw 'old', people may find difficulties to do because the 'old' is nonobjective and does not present clear visual 'prototype'. As illustrated in the upper row of Fig. 1, the dogs with different attributes present the similar visual feature. On the contrary, as shown in the lower row, the visual feature of attribute 'old' varies dramatically for different object classes. As a result, the discriminative model is not such successful for attribute recognition as that for object recognition, which further results in the low accuracy of attribute-object pair recognition. In fact, attribute is highly dependent on object. For example, when we teach a baby to recognize the attribute 'old', we often use instances like 'old book', 'old bike', and 'old dog' to show how 'old' looks like. Therefore, to better recognize attribute-object pair, we should explore the inner relation of the attribute and object instead of composing the discriminative models that are separately trained for the object and attribute.

Figure 1: The upper row shows the same object 'dog' with the different attributes. Though dogs can be small, big, wet, or wrinkled, they present similar visual properties. The lower row shows the same attribute 'old' with the different objects, we can observe that the visual properties significantly differ from each other.

*Yang Liu and Zhixiong Nan are co-first authors.

Realizing this issue, some works like (Chen and Grauman 2014) have tried to process the attribute and object as a whole to explore their inner relations. However, many of them still employ discriminative models to tackle the problem, resulting in poor performance for recognizing unseen attribute-object pairs. The major reason is that the individual property of the attribute and object are not learned. For example, there are pairs like 'old book' and 'small dog' in the training set, so the model are well fitted to these pairs. However, the model fails to learn the concept of individual attribute and object like 'old', 'book', 'small', and 'dog', thus can hardly generalize to unseen pairs like 'old dog' or 'small book'.

Summing up the above, to recognize unseen attribute-object pairs, we should design a model that should consider not only the individual property of the attribute and object but also the inner relation between them. To this end, in this paper, we propose an encoder-decoder generative model bridging visual and linguistic features in a unified end-to-end network. We first obtain the visual feature of images using state-of-the-art deep neural network, and the linguistic feature by extracting the semantic word embedding vectors of object and attribute label. To explore the inner relation of the attribute and object, inspired by the idea of ZSL (Zero-Shot Learning), we project the visual feature and the linguistic feature into a latent space where the attribute and object are processed as a whole. During the projection, to preserve the individual property of the attribute and object, the original visual and linguistic features of the attribute and object are projected by different functions. In the latent space, in order to minimize the 'distance' between the visual feature and linguistic feature, we have exploited several loss functions to penalize the dissimilarity between them. In addition, we propose the decoding loss which has been proved crucial to generalize better to unseen pairs because it can find the natural and intrinsic feature representations.

In the experiments, we have compared with four state-of-the-art methods on two challenging datasets, MIT-States (Isola, Lim, and Adelson 2015) and UT-Zappos50K (Yu and Grauman 2014). The experiments show that our method outperforms other methods significantly. We also performed some ablation experiments to study the effect of individual loss function, the influence of visual feature extractor, and the interdependence of the attribute and object, from which we draw some important conclusions.

our contributions in this paper are as follows:

- We propose a generative model to recognize unseen attribute-object pairs instead of composing multiple classifiers.
- Our model combines the visual and linguistic information in the same latent space, which is significant for exploring the inner relation of the attribute and object.
- We apply the encoder-decoder mechanism to the problem of attribute-object pair recognition.

Related Work

Object and attribute In recent years, object, attribute, and attribute-object composition recognition have been intensively studied in both image and video domains (Laffont et al. 2014; Isola, Lim, and Adelson 2015; Su et al. 2016; Wu et al. 2017; Misra, Gupta, and Hebert 2017; Liu, Wei, and Zhu 2017). Attribute recognition is a basic problem, the typical method for attribute recognition is similar to that for object classification, training discriminative models using attribute-labeled samples (Parikh and Grauman 2011; Patterson and Hays 2012; Scheirer et al. 2012; Singh and Lee 2016; Lu et al. 2017). Attribute-object pair recognition is a more challenging problem. Some conventional methods often use a classifier or compose multiple classifiers to tackle the problem (Chen and Grauman 2014; Misra, Gupta, and Hebert 2017). Some other studies assume object-attribute relationship is known and datasets are simple that only contain one or few dominant object categories (Wang and Mori 2010; Mahajan, Sellamanickam, and Nair 2011; Wang and Ji 2013). To make the model applicable to complex datasets that cover various object and attribute classes, some good models are proposed (Wang et al. 2013; Kulkarni et al. 2013). However, they are suffering from 'domain shift' problem - the test data distribution differs from that estimated by the training data, leading to the low performance on testing data. To overcome this problem, the work (Nagarajan and Grauman 2018) proposes to take the attributer as the operator and attribute-object pair as a vector that is transformed by this operator, then this transformed vector is compared with CNN visual feature to recognize unseen pairs. In this paper, we propose a generative model with encoder-decoder mechanism which is significant for exploring the intrinsic feature representation, thus can better transfer the concept of object and attribute from training set to the testing set.

Zero-shot learning The goal of zero-shot learning (ZSL) is to recognize unseen/new objects by utilizing their auxiliary information such as attribute or text description. One major method for zero-shot learning is first mapping the input into a semantic space where the auxiliary information like attributes of unseen objects are known, then finding the object whose auxiliary feature is 'closest' to the input feature (Lampert, Nickisch, and Harmeling 2014; Akata et al. 2015). Another major method learns a latent space that the input and the auxiliary feature of unseen objects are simultaneously projected into (Changpinyo et al. 2016; Wang et al. 2016), and the most likely unseen object is recognized by measuring the 'distance' between input feature and auxiliary feature in the latent space. Some other methods predict unseen objects using the classifier that is composed by seen object classifiers (Norouzi et al. 2013; Changpinyo et al. 2016). Recently, semantic autoencoder (SAE) has been proposed for zero-shot learning, considering both projection from input space to semantic space and the reverse, which demonstrates to be a simple but effective approach (Kodirov, Xiang, and Gong 2017). Several other works like (Wang et al. 2018) have explored more general generative methods using highly nonlinear model instead of linear regression from the latent space to the input space (Verma et al. 2018). Inspired by these works, in this paper, we project both visual and linguistic features into the same

latent space where the most likely attribute-object pair is selected with the least loss calculated by our self-defined loss function.

Vision and language combination With the rapid development of vision and language, vision and language combination has been studied to tackle many problems. For example, as mentioned above, many ZSL models take linguistic text description (Lei Ba et al. 2015; Elhoseiny, Saleh, and Elgammal 2013) as auxiliary information for unseen object recognition. However, text description annotation is 'expensive', especially for large scale datasets. Therefore, it is intuitive to utilize linguistic word embedding as auxiliary information because all words can be encoded as vectors with the pre-trained model (Socher et al. 2013; Frome et al. 2013). In this paper, we represent object and attribute as linguistic word embedding vectors. Different from one-hot vectors, word embedding vectors imply the semantic similarity of their corresponding words. In another word, semantically similar attribute and object will create similar word embedding vectors, which is significant for learning the inner relation of the attribute and object.

Approach

In this paper, we are studying the problem of identifying the attribute-object pair of the given image. For example, given an image as shown in Fig. 2, our task is to output the attribute-object pair 'wrinkled dog'. It is challenging for two reasons: 1) we are recognizing unseen attribute-object pairs that are not included in training data, and 2) this is fine-grained recognition problem and the number of possible attribute-object pairs is large.

Overview

Given an image I with the attribute label y_a and object label y_o, our goal is to correctly choose its attribute-object label from the set $Z = \{z_i^{\mathcal{L}}\}_{i=1}^{N}$ that contains all possible N attribute-object pairs. To realize this goal, the intuitive idea is combining the classifiers that are separately trained for the attribute and object. For example, to recognize unseen attribute-object pair 'small dog', some studies first learn the concept of 'small' from images like 'small cat', 'small horse', and other small objects as well as the concept of 'dog' from images like 'wrinkled dog', 'big dog' and other dogs using training set, and then combine the separate classifiers to recognize unseen attribute-object pair 'small dog'. However, as we have discussed previously that the attribute does not have clear visual 'prototype' and is highly dependent on the object. Therefore, we propose a generative model that combine the visual and linguistic information in the same latent space where the attribute and object are processed as a whole. As shown in Fig. 2, We first use deep neural networks (Simonyan and Zisserman 2014; He et al. 2016) to extract the visual feature of I in visual space $\mathcal{V}$ and denote it as $x^{\mathcal{V}}$, which is then projected into latent space $\mathcal{L}$ as $x^{\mathcal{L}}$. For all possible pairs, we extract their attribute vector $a^{\mathcal{S}}$ and object vector $o^{\mathcal{S}}$, and then project them into latent space $\mathcal{L}$ where they are merged as linguistic attribute-object pair feature $z^{\mathcal{L}}$. In the training stage, we are trying to learn the projection from $\mathcal{V}$ and $\mathcal{S}$ to $\mathcal{L}$ by minimizing the 'distance' between the visual feature $x^{\mathcal{L}}$ and its corresponding linguistic feature $z^{\mathcal{L}}$ in the latent space. In the testing stage, we predict the attribute-object label of $z^{\mathcal{L}}$ that is the closest to the $x^{\mathcal{L}}$. To this end, we need to consider two crucial problems: 1) how is original visual and linguistic data transitioned between different spaces to obtain $x^{\mathcal{L}}$ and $z^{\mathcal{L}}$, and 2) how to design the loss functions to minimize their 'distance'. We will tackle these two problems in the following sub-sections.

Data transitions

Fig. 2 signals the data transition process of our method. In the figure, the red circles represent the visual data flow, while green and blue waves represent the linguistic data flow. For visual data flow, the visual feature $\mathbf{x}^{\mathcal{V}}$ in visual space $\mathcal{V}$ is projected into latent space $\mathcal{L}$ as two flows representing visual attribute feature $x_a^{\mathcal{L}}$ and visual object feature $x_o^{\mathcal{L}}$ respectively. $x_a^{\mathcal{L}}$ and $x_o^{\mathcal{L}}$ are then merged as visual attribute-object pair feature $x^{\mathcal{L}}$. To obtain reconstruction of the original visual feature, $x^{\mathcal{L}}$ is re-mapped back to visual space $\mathcal{V}$, the re-mapped feature is denoted as $\hat{x}^{\mathcal{V}}$. For linguistic data flow, the linguistic attribute feature $a^{\mathcal{S}}$ and linguistic object feature $o^{\mathcal{S}}$ in word embedding space $\mathcal{S}$ are projected into the latent space as $z_a^{\mathcal{L}}$ and $z_o^{\mathcal{L}}$, respectively. Then $z_a^{\mathcal{L}}$ and $z_o^{\mathcal{L}}$ are merged as linguistic attribute-object pair feature $z^{\mathcal{L}}$.

For the projections of linguistic attribute and object features from $\mathcal{S}$ to $\mathcal{L}$, we define two projection functions $F^a_{\mathcal{S}\to\mathcal{L}}(\cdot)$ and $F^o_{\mathcal{S}\to\mathcal{L}}(\cdot)$ as linguistic encoders:

$$z_a^{\mathcal{L}} = F^a_{\mathcal{S}\to\mathcal{L}}(a^{\mathcal{S}}) \tag{1}$$

$$z_o^{\mathcal{L}} = F^o_{\mathcal{S}\to\mathcal{L}}(o^{\mathcal{S}}) \tag{2}$$

Here, we define different projection functions for the attribute and object because object (a noun in most cases) and attribute (an adjective to describe the noun in most cases) are two different kinds of instances and should be processed differently to explore their potential properties. The experiment results validate the effectiveness of this separate processing, the details can be found in the experiment section.

Based on the above two transitions, we can get the attribute feature and object feature individually. To explore their inner relation, we add them together:

$$z^{\mathcal{L}} = z_a^{\mathcal{L}} + z_o^{\mathcal{L}} \tag{3}$$

For visual feature in $\mathcal{V}$, we define two project functions $F^a_{\mathcal{V}\to\mathcal{L}}(\cdot)$ and $F^o_{\mathcal{V}\to\mathcal{L}}(\cdot)$ which project $x^{\mathcal{V}}$ to its attribute feature $x_a^{\mathcal{L}}$ and object feature $x_o^{\mathcal{L}}$.

$$x_a^{\mathcal{L}} = F^a_{\mathcal{V}\to\mathcal{L}}(x^{\mathcal{V}}) \tag{4}$$

$$x_o^{\mathcal{L}} = F^o_{\mathcal{V}\to\mathcal{L}}(x^{\mathcal{V}}) \tag{5}$$

With the same manner as we combine latent attribute feature and object feature from language. We get $x^{\mathcal{L}}$ by:

$$x^{\mathcal{L}} = x_a^{\mathcal{L}} + x_o^{\mathcal{L}} \tag{6}$$

Our decoder to re-map the visual feature from latent space $\mathcal{L}$ to original visual space $\mathcal{V}$ is defined as:

$$\hat{x}^{\mathcal{V}} = F^{pair}_{\mathcal{L}\to\mathcal{V}}(x^{\mathcal{L}}) \tag{7}$$

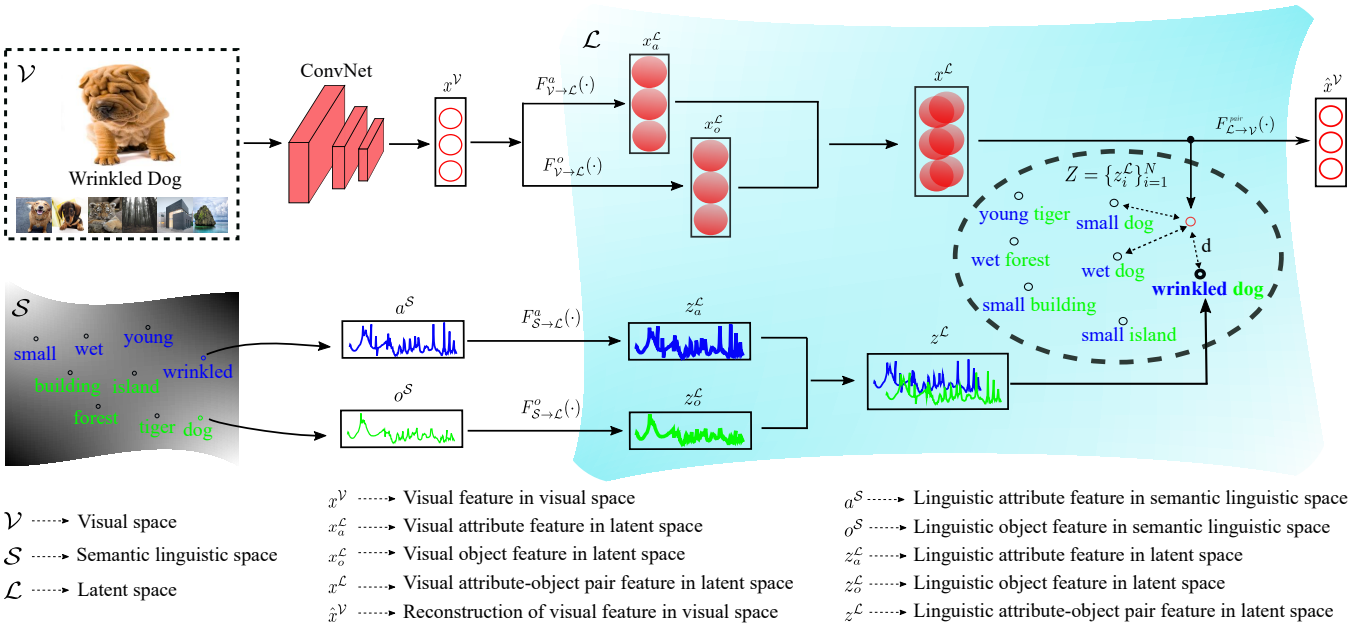

Figure 2: Given an image with the attribute-object pair label 'wrinkled dog', our goal is to correctly recognize this label. The challenge is that there are many possible pairs. As shown in the black ellipse in latent space $\mathcal{L}$, the pair may be 'young tiger', 'small dog', or others. To find the correct one, we first extract the visual feature $x^{\mathcal{V}}$ of the given image in visual space $\mathcal{V}$ and project $x^{\mathcal{V}}$ into latent space $\mathcal{L}$. After some processing, we finally obtain $x^{\mathcal{L}}$ that represents the visual feature of the given image in $\mathcal{L}$. At the same time, for all possible pairs, we extract their linguistic representations $Z = \{z_i^{\mathcal{L}}\}_{i=1}^N$ in latent space $\mathcal{L}$. Each element $z_i^{\mathcal{L}}$ in Z corresponds to one possible pair, which is obtained by mapping the corresponding word embedding vectors $a^{\mathcal{S}}$ and $o^{\mathcal{S}}$ from semantic linguistic space $\mathcal{S}$ to latent space $\mathcal{L}$. Finally, we take the label of $z_i^{\mathcal{L}}$ with closest distance to $x^{\mathcal{L}}$ as recognition result.

Loss functions

Encoding loss The goal of encoding loss is to minimize the 'distance' between visual attribute-object pair feature $x^{\mathcal{L}}$ and linguistic attribute-object pair feature $z^{\mathcal{L}}$. There are multiple ways to define the distance in the latent space. The most intuitive and common one is L2 distance. However, when using L2 distance, the value of visual and linguistic features in latent space tend to be extremely small during optimization which leads to poor performance. Therefore, we define the encoding loss as cosine similarity between $x^{\mathcal{L}}$ and $z^{\mathcal{L}}$:

$$L_{en} = dist(x^{\mathcal{L}}, z^{\mathcal{L}}) = 1 - \frac{< x^{\mathcal{L}}, z^{\mathcal{L}} >}{\|x^{\mathcal{L}}\|_2 \|z^{\mathcal{L}}\|_2}, \tag{8}$$

From geometrical aspect, we measure the angle between two vectors, and encourage them to have the same direction. The reason we do not consider the cosine similarity between $x_o^{\mathcal{L}}$ and $z_o^{\mathcal{L}}$ as well as the similarity between $x_a^{\mathcal{L}}$ and $z_a^{\mathcal{L}}$ is that individual attribute or object feature includes limited information for attribute-object pair estimation. In another word, we treat the attribute and object in the latent space as a whole to explore their inner relation.

Triplet loss The encoding loss defined in the Eq. 8 encourages the visual feature to be close to the linguistic feature of the indexed attribute-object pair, but doesn't consider that the visual feature should be far from the linguistic feature of other attribute-object pairs. Therefore, we add an extra loss called triplet loss, which impels the distance between $x^{\mathcal{L}}$ and $z^{\mathcal{L}}$ to be smaller than the distance between $x^{\mathcal{L}}$ and other linguistic attribute-object features $\tilde{z}^{\mathcal{L}}$ by a margin K:

$$L_{tri} = max\left(0, \frac{dis(x^{\mathcal{L}}, z^{\mathcal{L}})}{dist(x^{\mathcal{L}}, \tilde{z}^{\mathcal{L}})} - K\right) \tag{9}$$

Decoding loss Inspired by recent zero-shot learning works using autoencoder (Kodirov, Xiang, and Gong 2017; Wang et al. 2018), to explore the intrinsic representation of the input image, we introduce the decoding loss which is defined as the L2 distance between original visual feature $x^{\mathcal{V}}$ and reconstructed visual feature $\hat{x}^{\mathcal{V}}$:

$$L_{de} = \|x^{\mathcal{V}} - \hat{x}^{\mathcal{V}}\|_2 \tag{10}$$

The decoding loss encourages $\hat{x}^{\mathcal{V}}$ to be same with $x^{\mathcal{V}}$ rather than minimizing the angle between $\hat{x}^{\mathcal{V}}$ and $x^{\mathcal{V}}$. Therefore, we utilize the L2 loss instead of cosine similarity. We did not apply the decoding loss to linguistic features. The reason is that one attribute-object class only corresponds to one linguistic vector, the number of attribute-object classes is too small to learn a reprojection function with a huge number of parameters.

Discriminative loss In the above, we have introduced the encoding loss to encourage the visual attribute-object pair

feature to be close to the indexed linguistic attribute-object pair feature. However, this may lead to that the dominance effect of attribute or object. In another word, either attribute or object tends to represent the whole pair. To avoid this imbalance case, we define the discriminative loss to encourage to preserve the information for attributes and objects. The discriminative loss consists of attribute discriminative loss and object discriminative loss:

$$L_{dis} = L_{dis,a} + L_{dis,o} \tag{11}$$

$L_{dis,a}$ and $L_{dis,o}$ are defined as

$$L_{dis,a} = h(x_a^{\mathcal{L}}, y_a), \;\; L_{dis,o} = h(x_o^{\mathcal{L}}, y_o)$$

where $h(\cdot)$ is a one fully connected layer network with cross-entropy loss.

The purpose of the discriminative loss is to stress the individual property of the attribute and object in visual domain. The experiments validate the effectiveness of the discriminative loss, the details can be found the in experiment section.

Learning and Inference

The purpose of learning is to estimate the parameters of data transition functions defined in Eq. 1-7. Let W be all parameters of functions involved in Eq. 1-7. Given a set of images with the attribute and object labels, the estimation of W is equal to minimize the losses defined in Eq. 8-11 where κ,α, β, γ are the weights for different losses:

$$W^* = \arg\min_W (\kappa L_{en} + \alpha L_{tri} + \beta L_{de} + \gamma L_{dis}) \tag{12}$$

During inference, when a new image arrives, the visual feature is extracted and then projected to latent space, producing the visual representation $x^{\mathcal{L}}$ in latent space. At the same time, the linguistic features of all N possible attribute-object pairs are also computed to obtain a set of linguistic representations $Z = \{z_i^{\mathcal{L}}\}_{i=1}^N$. We compute the cosine similarity between $x^{\mathcal{L}}$ and every $z_i^{\mathcal{L}}$, and select the indexed attribute-object pair label of the most similar $z_i^{\mathcal{L}}$ as recognition result. In another word, we predict the attribute-object pair label by choosing the $z_i^{\mathcal{L}}$ with the least encoding loss.

Experiments

Setup

Datasets Two datasets, MIT-States (Isola, Lim, and Adelson 2015) and UT-Zappos50K (Yu and Grauman 2014), are used for evaluation. The MIT-States is a big dataset with 63,440 images. Each image is annotated with an attribute-object pair such as 'small bus'. It covers 245 object classes and 115 attribute classes. However, it does not have 245×115 attribute-object pairs as labels because not all pairs are meaningful in real world. Following the same setting as in (Misra, Gupta, and Hebert 2017) and (Nagarajan and Grauman 2018), 1,262 attribute-object pairs are used for training and 700 pairs for test. The training pairs and testing pairs are non-overlapping. UT-Zappos50k is a fine-grained shoe dataset with 50,025 images. Following the same setting as in (Nagarajan and Grauman 2018) We use 83 attribute-object pairs for training and 33 pairs for testing. The training pairs and testing pairs are also non-overlapping.

Baselines and metric We widely compare with four baseline methods, three of them are recently proposed state-of-the-art methods. ANALOG (Chen and Grauman 2014) predicts unseen attribute-object pairs using a sparse set of seen object-specific attribute classifiers. SAE (Kodirov, Xiang, and Gong 2017) predicts unseen pairs by projecting the input feature in a semantic space where the auxiliary information of unseen pairs is known. REDWINE (Misra, Gupta, and Hebert 2017) predicts unseen attribute-object pairs by composing existing attribute and object classifiers. ATT-OPERATOR (Nagarajan and Grauman 2018) represents the attribute-object pair as the object vector transformed by attribute operator, the transformed vector is compared with CNN visual feature for unseen pair recognition. We use the top-1 accuracy on testing images as evaluation metric.

Implementation details We extract 512 dimension visual feature of the image using ResNet-18 (He et al. 2016) pre-trained on ImageNet (Russakovsky et al. 2015). The network is not fine-tuned on MIT-States or UT-Zappos50K dataset. We extract 300 dimension linguistic feature for object and attribute using pre-trained word vectors (Pennington, Socher, and Manning 2014), some not-included words are substituted by synonyms. All these features are mapped into a 1024 dimension latent space. K in Eq. 9 is a parameter that controls the margin, and is set to 0.9 in our experiment. κ,α, β, γ in Eq. 12 are with the ratio of $1 : 0.2 : 2 : 2$ for Mit-States dataset and $1 : 0.2 : 0.5 : 2$ for UT-Zappos50K dataset. We implement our end to end neural network with MXNet (Chen et al. 2015). For all the projection functions, we implement each as one fully connected layer. For the projections from visual space to latent space, to resolve overfitting problem, we add dropout layers after each fully connected layer with dropout ratio as 0.3. We use ADAM as our optimizer with the initial learning rate as 0.0001, which decays by 0.9 every two epochs. At every iteration we feed the mini-batch to the network with the batch size as 128.

Quantitative results

As shown in Tab 1, our method outperform all recently proposed methods, achieving 25.4% and 4.5% improvement over the second best methods respectively on MIT-States and UT-Zappos50k datasets. Our method outperforms others for two reasons: 1) we introduce the encoder-decoder mechanism that enables to learn the general and intrinsic representation of attributes-object pair, and 2) our model considers not only the individual property of the attribute and object but also the inner relation between them.

The average accuracy on UT-Zappos50k is higher than that on MIT-States. This mainly results from the difference of data complexity. Images in UT-Zappos50k have single white background as shown in Fig. 3, and few attribute and object classes, while images in MIT-States cover a variety of backgrounds, object classes, and attribute classes. In addition, only 33 attribute-object pairs are used for testing in UT-Zappos50k, while 700 pairs are used in MIT-States.

Methods	MIT-States(%)	UT-Zappos(%)
CHANCE	0.14	3.0
ANALOG(Chen and Grauman 2014)	1.4	18.3
SAE (Kodirov, Xiang, and Gong 2017)	14.0	31.0
REDWINE (Misra, Gupta, and Hebert 2017)	12.5	40.3
ATTOPERATOR (Nagarajan and Grauman 2018)	14.2	46.2
Ours	**17.8**	**48.3**

Table 1: Top-1 accuracy of methods tested on the MIT-States dataset and UT-Zappos50k dataset. For fair comparison, all methods use the same visual feature extracted with ResNet-18.

Figure 3: Qualitative results on two datasets. For each dataset, the fifth column shows some false recognitions and other columns show the true recognitions.

Qualitative results

Fig. 3 shows some qualitative results on MIT-States dataset and UT-Zappos50K dataset. Columns with the green mark show some true recognitions. We can observe that some samples with extremely abstract attribute-object pairs are correctly recognized. For example, some images like 'wet forest' and 'dry forest' are correctly recognized. However, the accuracy is low. On one hand, forest is an abstract object sharing similar properties with objects like tree, bush, plant, and jungle. On the other hand, wet and dry are abstract attributes sharing the similar properties with attributes like damp, verdant, mossy, and barren. Factually for some 'wet forest' testing images, our model has recognized them as 'damp bush', 'mossy jungle', 'verdant plant' and other similar compositions. This demonstrates that our model has learned some macro concepts for attribute-object pairs, but some micro concepts have not been precisely distinguished.

Columns with the red mark shows some false recognitions. We can observe that some images with the label of 'scratched phone' and 'broken camera' are wrongly recognized as 'broken phone', and pairs like 'synthetic ankle-boot' are wrongly recognized as 'rubber ankle-boot'. One main reason is that some attributes are similar and some objects present similar appearance. For example, the attribute of 'scratched' is similar with 'broken', the appearance of some cameras is similar to that of phone, and some 'synthetic' boots also present the attribute of 'rubber'.

Item	MIT-States		UT-Zappos	
	Top1 (%)	Top5(%)	Top1 (%)	Top5(%)
Att	15.1	38.9	18.4	76.8
Att (pair)	**25.1**	**55.3**	**52.0**	**92.7**
Obj	27.7	**56.4**	68.1	**96.7**
Obj (pair)	**29.9**	51.6	**77.3**	93.9

Table 2: 'Att' and 'Obj' correspond to the accuracies that are trained using the original visual feature, while 'Att (pair)' and 'Obj (pair)' correspond to the accuracies that are extracted from our attribute-object pair recognition results.

Attribute and object relation

To validate the relation of the attribute and object we have learned, we designed an experiment that measures the attribute and object recognition accuracy under two conditions: 1) not considering the relation of the attribute and object, as 'Att' and 'Obj' shown in Tab. 2, and 2) considering the relation of the attribute and object in both visual and linguistic domains, as 'Att (pair)' and 'Obj (pair)' shown in Tab. 2. Actually, 'Att (pair)' and 'Obj (pair)' correspond to the accuracy that are extracted from our attribute-object pair recognition results. While for 'Att' and 'Obj', we have separately trained a 2-layer MLP model to recognize attribute and object category. We can observe from Tab. 2 that attribute recognition accuracy of our model is always higher than that does not consider relation of the attribute and object, which validates our claim that attribute is highly dependent on object.

Ablation study

We design two experiments. One is to study the importance of different loss functions, the other is to study the effect of different visual features.

For the detail analysis of different loss functions, we report the accuracy corresponding to different kinds of loss function compositions on both MIT-States and UT-Zappos50K datasets as shown in Tab. 3. If only encoding loss is used, the accuracy is 3.6% on the MIT-States and 37.8% on the UT-Zappos50K. If we add triplet loss (+tri) to encoding loss, we obtain significant performance improvement on both datasets, which validate our claim that we should not only encourage the visual feature to be close to the linguistic feature of its indexed attribute-object pair but also should let it to be away from the linguistic features of the other attribute-object pairs. If we add discriminative loss

Loss	MIT-States(%)	UT-Zappos(%)
en	3.6	37.8
+tri	11.2	45.5
+dis	15.3	37.9
+de	17.2	41.3
+tri+dis	15.7	41.4
+tri+de	17.4	46.7
+dis+de	17.5	43.5
+de+dis+tri*	16.5	47.5
+de+dis+tri	**17.8**	**48.3**

Table 3: Accuracy for loss function ablation study. 'en', 'tri', 'dis', and 'de' represent the encoding loss, triplet loss, discriminative loss, and decoding loss, respectively. *means sharing parameters for $F^a_{\mathcal{S}\to\mathcal{L}}(\cdot)$ and $F^o_{\mathcal{S}\to\mathcal{L}}(\cdot)$

(+dis) to encoding loss, we obtain significant performance improvement on MIT-States but slight improvement on UT-Zappos50K. On one hand, we can conclude that the discriminative loss allows to learn better visual attribute-object representation by stressing individual attribute and object property. On the other hand, it is based on certain condition. The MIT-States dataset is complex and has many object and attribute classes, while the UT-Zappos50K is relatively simple that the visual features already have good representations, so adding discriminative loss only contributes slightly. If we add decoding loss (+de), we obtain impressive performance improvement on the MIT-States, even better than adding both triplet loss and discriminative loss (+tri+dis), which demonstrates that encoder-decoder mechanism can mine the essential representation for attribute-object pair. On the UT-Zappos50K, the decoding loss is also helpful.

From the Tab. 3 we can observe that the performance is basically increasing when adding more loss functions. Though in some cases adding more losses lead to worse results, the best performance is achieved when using all loss functions (+de+dis+tri), from which we can conclude that the four loss functions are complementary to each other. In the Tab. 3, (+de+dis+tri*) corresponds to the accuracy when we impose the constraint that the attribute and object are processed by the same projection function (sharing the parameters in Eq. 1 and Eq. 2). We can observe that the accuracy of (+de+dis+tri*) is lower than that of (+de+dis+tri) on both datasets, from which we can draw another important conclusion that object and attribute are two different kinds of instances and should be processed differently to better explore their individual properties. .

In Tab. 4, we report the accuracy corresponding to different kinds of visual feature extractors. We tested two kinds of network architectures, VGG (Simonyan and Zisserman 2014) and ResNet (He et al. 2016). We can observe from table that the visual feature significantly affects the final performance. VGG-19 presents similar performance with VGG-16. ResNet behaves better than VGG, and basically achieves higher accuracy with deeper architectures.

Network	MIT-States(%)	UT-Zappos(%)
VGG-16	15.4	40.7
VGG-19	15.3	40.8
ResNet-18	17.8	48.3
ResNet-50	19.7	**52.0**
ResNet-101	**20.0**	51.9

Table 4: Accuracy of our method with different visual feature extractors.

Conclusion

In this paper, to recognize the unseen attribute-object pair of a given image, we propose an encoder-decoder generative model to bridge visual and linguistic features in a unified end-to-end network. By comparing our method with several state-of-the-art methods on two datasets, we reach the conclusion that 1) the generative model is more competitive than discriminative models to recognize unseen classes, 2) the encoder-decoder mechanism is crucial for learning intrinsic feature representations, and 3) an appropriate model should consider not only the individual property of the attribute and object but also the inner relation between them.

Acknowledgment

This research is supported by DARPA XAI grant N66001-17-2-4029, ONR MURI grant N00014-16-1-2007, and ARO grant W911NF-18-1-0296 from USA. This research is also supported in part by the National Natural Foundation of China under Grand 61773312.

References

Akata, Z.; Reed, S.; Walter, D.; Lee, H.; and Schiele, B. 2015. Evaluation of output embeddings for fine-grained image classification. *IEEE Conference on Computer Vision and Pattern Recognition*.

Changpinyo, S.; Chao, W.-L.; Gong, B.; and Sha, F. 2016. Synthesized classifiers for zero-shot learning. *IEEE Conference on Computer Vision and Pattern Recognition*.

Chen, C.-Y., and Grauman, K. 2014. Inferring analogous attributes. *IEEE Conference on Computer Vision and Pattern Recognition*.

Chen, T.; Li, M.; Li, Y.; Lin, M.; Wang, N.; Wang, M.; Xiao, T.; Xu, B.; Zhang, C.; and Zhang, Z. 2015. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint arXiv:1512.01274*.

Elhoseiny, M.; Saleh, B.; and Elgammal, A. 2013. Write a classifier: Zero-shot learning using purely textual descriptions. *IEEE International Conference on Computer Vision*.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Mikolov, T.; et al. 2013. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and pattern recognition*.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. *IEEE Conference on Computer Vision and Pattern Recognition*.

Isola, P.; Lim, J. J.; and Adelson, E. H. 2015. Discovering states and transformations in image collections. *IEEE Conference on Computer Vision and pattern recognition*.

Kodirov, E.; Xiang, T.; and Gong, S. 2017. Semantic autoencoder for zero-shot learning. *arXiv preprint arXiv:1704.08345*.

Kulkarni, G.; Premraj, V.; Ordonez, V.; Dhar, S.; Li, S.; Choi, Y.; Berg, A. C.; and Berg, T. L. 2013. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kumar, N.; Belhumeur, P.; and Nayar, S. 2008. Facetracer: A search engine for large collections of images with faces. *European Conference on Computer Vision*.

Laffont, P.-Y.; Ren, Z.; Tao, X.; Qian, C.; and Hays, J. 2014. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on Graphics*.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2014. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Lei Ba, J.; Swersky, K.; Fidler, S.; et al. 2015. Predicting deep zero-shot convolutional neural networks using textual descriptions. *IEEE International Conference on Computer Vision*.

Liu, Y.; Wei, P.; and Zhu, S.-C. 2017. Jointly recognizing object fluents and tasks in egocentric videos. *IEEE International Conference on Computer Vision*.

Lu, Y.; Kumar, A.; Zhai, S.; Cheng, Y.; Javidi, T.; and Feris, R. S. 2017. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. *IEEE Conference on Computer Vision and Pattern Recognition*.

Mahajan, D.; Sellamanickam, S.; and Nair, V. 2011. A joint learning framework for attribute models and object descriptions. *IEEE International Conference on Computer Vision*.

Misra, I.; Gupta, A.; and Hebert, M. 2017. From red wine to red tomato: Composition with context. *IEEE Conference on Computer Vision and Pattern Recognition*.

Nagarajan, T., and Grauman, K. 2018. Attributes as operators. *European Conference on Computer Vision*.

Norouzi, M.; Mikolov, T.; Bengio, S.; Singer, Y.; Shlens, J.; Frome, A.; Corrado, G. S.; and Dean, J. 2013. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*.

Parikh, D., and Grauman, K. 2011. Relative attributes. *IEEE International Conference on Computer Vision*.

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

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. *Empirical Methods in Natural Language Processing*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*.

Scheirer, W. J.; Kumar, N.; Belhumeur, P. N.; and Boult, T. E. 2012. Multi-attribute spaces: Calibration for attribute fusion and similarity search. *IEEE Conference on Computer Vision and Pattern Recognition*.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Singh, K. K., and Lee, Y. J. 2016. End-to-end localization and ranking for relative attributes. *European Conference on Computer Vision*.

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

Su, C.; Zhang, S.; Xing, J.; Gao, W.; and Tian, Q. 2016. Deep attributes driven multi-camera person re-identification. *European Conference on Computer Vision*.

Verma, V. K.; Arora, G.; Mishra, A.; and Rai, P. 2018. Generalized zero-shot learning via synthesized examples. *IEEE Conference on Computer Vision and Pattern Recognition*.

Wang, X., and Ji, Q. 2013. A unified probabilistic approach modeling relationships between attributes and objects. *IEEE International Conference on Computer Vision*.

Wang, Y., and Mori, G. 2010. A discriminative latent model of object classes and attributes. *European Conference on Computer Vision*.

Wang, S.; Joo, J.; Wang, Y.; and Zhu, S.-C. 2013. Weakly supervised learning for attribute localization in outdoor scenes. *IEEE Conference on Computer Vision and Pattern Recognition*.

Wang, W.; Chen, C.; Chen, W.; Rai, P.; and Carin, L. 2016. Deep metric learning with data summarization. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*.

Wang, W.; Pu, Y.; Verma, V. K.; Fan, K.; Zhang, Y.; Chen, C.; Rai, P.; and Carin, L. 2018. Zero-shot learning via class-conditioned deep generative models. *AAAI Conference on Artificial Intelligence*.

Wu, Q.; Shen, C.; Wang, P.; Dick, A.; and van den Hengel, A. 2017. Image captioning and visual question answering based on attributes and external knowledge. *IEEE transactions on pattern analysis and machine intelligence*.

Yan, X.; Yang, J.; Sohn, K.; and Lee, H. 2016. Attribute2image: Conditional image generation from visual attributes. *European Conference on Computer Vision*.

Yu, A., and Grauman, K. 2014. Fine-grained visual comparisons with local learning. *IEEE Conference on Computer Vision and Pattern Recognition*.