Selective Zero-Shot Classification with Augmented Attributes

Jie Song[1], Chengchao Shen[1], Jie Lei[1], An-Xiang Zeng[2], Kairi Ou[2],
Dacheng Tao[3], and Mingli Song[1](✉)

[1] College of Computer Science and Technology, Zhejiang University,
Hangzhou, China
{sjie,chengchaoshen,ljaylei,brooksong}@zju.edu.cn
[2] Alibaba Group, Hangzhou, China
{renzhong,suzhe.okr}@taobao.com
[3] UBTECH Sydney AI Centre, SIT, FEIT,
University of Sydney, Camperdown, Australia
dacheng.tao@sydney.edu.au

Abstract. In this paper, we introduce a selective zero-shot classification problem: how can the classifier avoid making dubious predictions? Existing attribute-based zero-shot classification methods are shown to work poorly in the selective classification scenario. We argue the under-complete human defined attribute vocabulary accounts for the poor performance. We propose a selective zero-shot classifier based on both the human defined and the automatically discovered residual attributes. The proposed classifier is constructed by firstly learning the defined and the residual attributes jointly. Then the predictions are conducted within the subspace of the defined attributes. Finally, the prediction confidence is measured by both the defined and the residual attributes. Experiments conducted on several benchmarks demonstrate that our classifier produces a superior performance to other methods under the risk-coverage trade-off metric.

Keywords: Zero-shot classification · Selective classification
Defined attributes · Residual attributes · Risk-coverage trade-off

1 Introduction

Zero-Shot Classification (ZSC) addresses the problem of recognizing images from novel categories, *i.e.*, those categories which are not seen during the training phase. It has attracted much attention [1–6] in the last decade due to its importance in real-world applications, where the data collection and annotation are both laboriously difficult. Existing ZSC methods usually assume that both the seen and the unseen categories share a common semantic space (*e.g.*, attributes [1,2]) where both the images and the class names can be projected. Under this assumption, the recognition of images from unseen categories can be achieved by the nearest neighbor search in the shared semantic space.

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11213, pp. 474–490, 2018.
https://doi.org/10.1007/978-3-030-01240-3_29

Although there is a large literature on ZSC, the prediction of existing zero-shot classifiers remains quite unreliable compared to that of the fully supervised classifiers. This limits their deployment in real-world applications, especially where mistakes may cause severe risks. For example, in autonomous driving, a wrong decision can result in traffic accidents. In clinical trials, a misdiagnosis may make the patient suffer from great pain and loss.

To reduce the risk of misclassifications, selective classification improves classification accuracy by rejecting examples that fall below a confidence threshold [7,8]. Motivated by this, in this paper we introduce a Selective Zero-Shot Classification (Selective ZSC) problem: the zero-shot classifier can abstain from predicting when it is uncertainty about its predictions. It requires that the classifier not only makes accurate predictions given images from unseen categories but also be self-aware. In other words, the classifier should be able to know when it is confident (or uncertain) about their predictions. The confidence is typically quantified by a confidence score function. Equipped with this ability, the classifier can leave the classification of images when it is uncertain about its predictions to the external domain expert (*e.g.,* drivers in autonomous driving, or doctors in clinical trials).

Selective classification is an old topic in machine learning field. However, we highlight its importance in the context of ZSC in threefold. Firstly, the predictions of zero-shot classifiers are not so accurate compared with those of fully supervised classifiers, which poses large difficulty in Selective ZSC. Secondly, it is shown in our experiments (in Sect. 6.3) that most existing zero-shot classifiers exhibit poor self-awareness. This results in their inferior performance in the settings of Selective ZSC. Lastly, albeit its great importance in real-world applications, selective classification remains under-studied in the field of ZSC.

Typically, existing ZSC methods rely on human defined attributes for novel class recognition. Attributes are a type of mid-level semantic properties of visual objects that can be shared across different object categories. Manually defined attributes are often those nameable properties such as color, shape, and texture. However, the discriminative properties for the classification task are often not exhaustively defined and sometimes hard to be described in a few words or some semantic concepts. Thus, the under-complete defined attribute vocabulary results in inferior performance of attribute-based ZSC methods. We call the residual discriminative but not defined properties *residual attributes*. To make safer predictions for zero-shot classification, we argue both the defined and the residual attributes should be exploited. These two types of attributes together are named *augmented attributes* in this paper.

We propose a much safer selective classifier for zero-shot recognition based on augmented attributes. The proposed classifier is constructed by firstly learning the augmented attributes. Motivated by [9,10], we formulate the attribute learning task as a dictionary learning problem. After the learning of the augmented attributes, the defined attributes can be directly utilized to accomplish traditional zero-shot recognitions. The confidence function thus can be defined within the subspace of defined attributes. The residual attributes, however, can not be directly exploited for classification because there are no associations between the residual attributes and the unseen categories. Instead of conducting direct predictions, we leverage the residual attributes to improve the self-awareness of the classifier constructed on defined attributes. Specifically, we define another confidence function based on the consistency between the defined and the residual attributes. Combining the confidence obtained on the augmented attributes and confidence produced within the defined attributes, the proposed selective classifier significantly outperforms other methods in extensive experiments.

To sum up, we made the following contributions: (1) we introduce the selective zero-shot classification problem, which is important yet under-studied; (2) we propose a selective zero-shot classifier, which leverages both the manually defined and the automatically discovered residual attributes for safer predictions; (3) we propose a solution to the learning of residual discriminative properties in addition to the manually defined attributes; (4) experiments demonstrate our method significantly outperforms existing state-of-the-art methods.

2 Related Work

2.1 Zero-Shot Learning

Typically, existing ZSC methods consist of two steps. The first step is an embedding process, which maps both the image representations and the class names to a shared embedding space. This step can also be viewed as a kind of multi-modality matching problem [11,12]. The second step is a recognition process, which is usually accomplished by some form of nearest neighbor searches in the shared space learned from the first step. Existing ZSC approaches mainly differ in the choices for the embedding model and the recognition model. For example, DAP [1] adopts probabilistic attribute classifiers for embedding and Bayes classifier for recognition. Devise [13], Attribute Label Embedding (ALE) [14], Simple ZSC [3] and Structured Joint Embedding (SJE) [4] adopt linear projection and inner product for embedding and recognition, respectively. However, they exploit different objective functions for optimization. Embedding Model (LatEm) [15] and Cross Model Transfer (CMT) [16] employ nonlinear projection for embedding to overcome the limitations of linear models. Different from above methods, Semantic Similarity Embedding (SSE) [17], Convex Combination of Semantic Embeddings (CONSE) [18] and Synthesized Classifiers (SYNC) [19] build the shared embedding space by expressing images and semantic class embeddings as a mixture of seen class proportions. For a more comprehensive review about ZSC, please refer to [5,20].

2.2 Defined Attributes and Latent Attributes

Attributes are usually defined as the explainable properties such as color, shape, and parts. With manually defined attributes as a shared semantic vocabulary, novel classes can be easily defined such that zero-shot recognition can be accomplished via the association between the defined attributes and the categories.

However, manually finding a discriminative and meaningful set of attributes can sometimes be difficult. The method for learning discriminative latent attributes has been exploited [9,21–24]. Tamara *et al.* [21] propose to automatically identify attributes vocabulary from text descriptions of images sampled from the Internet. Viktoriia *et al.* [22] propose to augment defined attributes with latent attributes to facilitate few-shot learning. Mohammad *et al.* [23] propose to discover attributes by trading off between predictability and discrimination. Felix *et al.* [24] propose to design attributes without concise semantic terms for visual recognition by incorporating both the category-separability and the learnability into the learning criteria. Peixi *et al.* [9] propose a dictionary learning model to decompose the dictionary space into three parts corresponding to defined, latent discriminative and latent background attributes. Different from these works, in this paper we augment the manually defined attributes with residual attributes to improve the self-awareness of zero-shot classifier.

2.3 Selective Classification

Safety issues have attracted much attention in the AI research community in the last several years. For example, Szegedy *et al.* [25] find that deep neural networks are easily fooled by adversarial examples. Following their work, many methods are proposed to construct more robust classifiers.

To reduce the risk of misclassifications, selective classification [7,8] improve classification accuracy by rejecting examples that fall below a confidence threshold. For different classifiers, the confidence scores can be defined in various ways. Most generative classification models are probabilistic, therefore they provide such confidence scores in nature. However, most discriminative models do not have direct access to the probability of their predictions [26]. Instead, related non-probabilistic scores are used as proxies, such as the margin in the SVM classifier and the softmax output or MC-Dropout [27] in deep neural networks. In this paper, we propose to exploit the residual attributes to compensate the limitations of defined attributes and make the classifier more self-aware.

3 Problem Formulation of Selective Zero-Shot Classification

We summarize some key notations used in this paper in Table 1 for reference.

Let $\mathcal{X}$ be the feature space (e.g., raw image data or feature vectors) and $\mathcal{Y}$ be a finite label set. Let $P_{\mathcal{X},\mathcal{Y}}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$. In a standard multi-class zero-shot classification problem, given training data $\mathbf{X}_s = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{N_s}]$ and corresponding defined attribute annotations $\mathbf{D}_s = [\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_{N_s}]$ and label annotations $\mathbf{y}_s = [y_1, y_2, ..., y_{N_s}]^T, y_i \in \mathcal{Y}_s$, the goal is to learn a classifier f: $\mathcal{X} \rightarrow \mathcal{Y}$. The classifier is usually used to recognize test data $\mathbf{X}_u = [\mathbf{x}_1^u, \mathbf{x}_2^u, ..., \mathbf{x}_{N_u}^u]$ from $\mathcal{Y}_u \subset \mathcal{Y}$ which is unseen during training, *i.e.*, $\mathcal{Y}_s \cap \mathcal{Y}_u = \emptyset$.

In the proposed Selective ZSC problem, the learner should output a selective classifier defined to be a pair (f, g), where f is a standard zero-shot classifier,

Table 1. Some key notations used in this paper. Some of them are also explained in the main text.

Notations	Definition
$\mathcal{Y}_s, \mathcal{Y}_u$	Seen label set and unseen label set
N_s, N_u	Number of seen (unseen) images, N_s $(N_u) \in \mathbb{N}^+$
K_o	Number of dimensions of the feature space, $K_o \in \mathbb{N}^+$
$\mathbf{x}_i$	An instance in the feature space, $\mathbf{x}_i \in \mathbb{R}^{K_o}$
$\mathbf{X}_s, \mathbf{X}_u$	Seen/Unseen image representations, $\mathbf{X}_s \in \mathbb{R}^{K_o \times N_s}, \mathbf{X}_u \in \mathbb{R}^{K_o \times N_u}$
$\mathbf{y}_s$	Label annotations for the training data $\mathbf{X}_s$, $\mathbf{y}_s \in \mathbb{R}^{N_s}$
K_d, K_r	Number of dimensions of the defined and the residual attribute space
$\mathbf{D}_s$	Defined attribute annotations $\mathbf{D}_s \in \mathbb{R}^{K_d \times N_s}$ for the training data $\mathbf{X}_s$
$\mathbf{D}_o$	Defined attribute annotations $\mathbf{D}_o \in \mathbb{R}^{K_d \times \lvert\mathcal{Y}_s\rvert}$ for the seen classes
$\mathbf{R}_o$	Residual attribute representations $\mathbf{R}_o \in \mathbb{R}^{K_r \times \lvert\mathcal{Y}_s\rvert}$ for the seen classes
$[\mathbf{d}_i; \mathbf{r}_i]$	Augmented attribute representation of $\mathbf{x}_i$. $\mathbf{d}_i$ is the defined attributes, and $\mathbf{r}_i$ is the residual attributes
$[\mathbf{d}^j; \mathbf{r}^j]$	Augmented attribute representation of class j
$\mathbf{s}_d, \mathbf{s}_r$	Similarity vectors from the defined/residual attributes, $\mathbf{s}_d, \mathbf{s}_r \in \mathbb{R}^{\lvert\mathcal{Y}_s\rvert}$

and $g : \mathcal{X} \rightarrow \{0, 1\}$ is a selection function which is usually defined as $g(\mathbf{x}) = \mathbb{1}\{conf(\mathbf{x}) > \tau\}$. $conf$ is a confidence function, τ is a confidence threshold, and $\mathbb{1}$ is an indicator function. Given a test sample $\mathbf{x}$,

$$(f, g)(\mathbf{x}) \triangleq \begin{cases} f(\mathbf{x}), & g(\mathbf{x}) = 1 \\ reject, & g(\mathbf{x}) = 0 \end{cases} \tag{1}$$

The selective zero-shot classifier abstains from prediction when $g(\mathbf{x}) = 0$. Its performance is usually evaluated by the risk-coverage curve [8,28]. More details about the evaluation metric can be found in Sect. 6.1.

4 The Proposed Selective Zero-Shot Classifier

In this section, we assume the model for augmented attributes has been learned and introduce our proposed selective classifier (f, g) based on the augmented attributes. Then in the next section, we introduce how the augmented attributes are learned.

Let $\mathcal{D}$ be the defined attribute space and $\mathcal{R}$ be the residual attribute space. For each $\mathbf{x} \in \mathcal{X}$, we can obtain its augmented attribute prediction $[\mathbf{d}; \mathbf{r}] \in \mathcal{DR}$ by the trained attribute model, where $\mathcal{DR} = \mathcal{D} \times \mathcal{R}$. In zero-shot learning, for each seen category $y_s \in \mathcal{Y}_s$, an attribute annotation $\mathbf{d}^{y_s}$ of the defined attributes is given. $\mathbf{D}_o \in \mathbb{R}^{K_d \times \lvert\mathcal{Y}_s\rvert}$ is the class-level attribute annotation matrix, where the i-th column vector denotes the defined attribute annotation for the i-th seen category. Since no annotations of residual attributes are provided for the

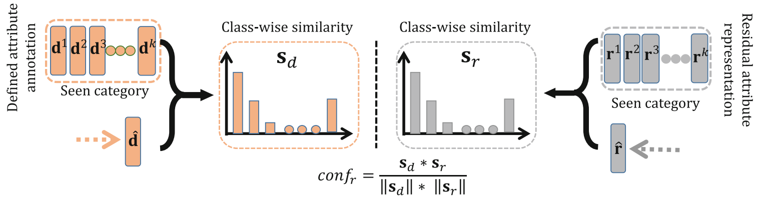
Defined attribute annotation
$\mathbf{d}^1$ $\mathbf{d}^2$ $\mathbf{d}^3$ $\mathbf{d}^k$
Seen category
$\hat{\mathbf{d}}$
Class-wise similarity
$\mathbf{s}_d$
Class-wise similarity
$\mathbf{s}_r$
$\mathbf{r}^1$ $\mathbf{r}^2$ $\mathbf{r}^3$ $\mathbf{r}^k$
Seen category
$\hat{\mathbf{r}}$
Residual attribute representation
$conf_r = \frac{\mathbf{s}_d * \mathbf{s}_r}{\|\mathbf{s}_d\| * \|\mathbf{s}_r\|}$

Fig. 1. The confidence defined with the aid of the residual attributes.

seen categories, we adopt the center of residual attribute predictions for each seen category as its residual attribute representation, denoted by $\mathbf{r}^{y_s}$. Let $\mathbf{R}_o \in \mathbb{R}^{K_r \times |\mathcal{Y}_s|}$ be the class-level residual attribute representation matrix. During the test phase, only the defined attributes are annotated for unseen categories ($\mathbf{d}^{y_u}$ for $y_u \in \mathcal{Y}_u$).

4.1 Zero-Shot Classifier f

The zero-shot classifier f is built on the defined attributes solely, as no annotations for residual attributes are provided. Given the defined attribute prediction $\hat{\mathbf{d}}$ of a test image, the classifier f is constructed by some form of nearest neighbor search

$$\hat{y} = \arg\max_{k \in \mathcal{Y}_u} sim(\hat{\mathbf{d}}, \mathbf{d}^k), \tag{2}$$

where sim is the similarity function. In fact, many ZSC approaches follow the above general formulation, even though they may differ in the concrete form of sim. In this paper, it is simply defined as the cosine similarity.

4.2 Confidence Function

With $sim(\cdot)$ defined within the subspace of the manually defined attributes, the prediction confidence can be defined as the similarity score:

$$conf_d = sim(\hat{\mathbf{d}}, \mathbf{d}^k). \tag{3}$$

However, as aforementioned, the defined attribute vocabulary alone is limited in its discriminative power. Thus the confidence score obtained within the defined attribute subspace is shortsighted. To tackle this issue, we propose to explore and exploit the residual attributes to overcome the shortcomings of the confidence produced by the defined attributes. Figure 1 illustrates the confidence score produced resorting to the residual attributes. Specifically, given a test image from an unseen class, we can obtain its augmented attribute presentation ($[\hat{\mathbf{d}}; \hat{\mathbf{r}}]$) by feeding the test image to the attribute prediction model. With this attribute presentation, two similarity vectors ($\mathbf{s}_d$, $\mathbf{s}_r$) can be computed: $\mathbf{s}_d$ for the defined attributes and $\mathbf{s}_r$ for the residual attributes. In these similarity vectors, the value of dimension k measures the similarity between the predicted attributes and attribute presentation of class k. We formulate the similarity vector learning task as a sparse coding problem:

$$\mathbf{s}_d = \arg\min_{\mathbf{s}} \left\{ \frac{\gamma}{2} \|\mathbf{s}\|^2 + \frac{1}{2} \left\| \hat{\mathbf{d}} - \mathbf{D}_o \mathbf{s} \right\|_F^2 \right\}, \tag{4}$$

$$\mathbf{s}_r = \arg\min_{\mathbf{s}} \left\{ \frac{\gamma}{2} \|\mathbf{s}\|^2 + \frac{1}{2} \|\hat{\mathbf{r}} - \mathbf{R}_o \mathbf{s}\|_F^2 \right\}. \tag{5}$$

Then the confidence score can then be defined as the consistency of these two vectors:

$$conf_r = sim(\mathbf{s}_d, \mathbf{s}_r). \tag{6}$$

The above confidence function is built on the intuition that the more consistent the defined and the residual attributes are, the less additional discriminative information the residual attributes provide for the current test image. Therefore, classification based on the defined attributes solely approximates classification based on the whole augmented attributes. Imagine that the residual attributes produce the same similarity vector as the defined attributes, then the residual attributes completely agree with the defined attribute on the prediction they made. However, if the residual attributes produce absolutely different similarity vector, then they do not reach a consensus. The defined attributes are short-sighted in this case and the produced prediction is more unreliable.

Combining the confidence function defined within the defined attribute subspace and that defined with the aid of residual attributes, the final confidence is

$$conf = (1 - \lambda) conf_d + \lambda conf_r, \tag{7}$$

where λ is a trade-off hyper-parameter which is set via cross-validation.

5 Augmented Attribute Learning

In this section, we introduce how the augmented attributes are learned. We formulate the augmented attribute learning task as a dictionary learning problem. The dictionary space is decomposed into two parts: (1) $\mathbf{Q}_d$ corresponding to the defined-attribute-correlated dictionary subspace part which is correlated to the defined attribute annotations and the class annotations, (2) $\mathbf{Q}_r$ corresponding to the residual attribute dictionary subspace which is correlated to the class annotations and thus also useful for the classification task. To learn the whole dictionary space, three criteria are incorporated: (1) the defined attributes alone should be able to accomplish the classification task as better as possible; (2) the residual attributes should complement the discriminative power of defined attributes for classification; (3) the residual attributes should not rediscover the patterns that exist in the defined attributes. With all the three criteria, the objective function is formulated as:

$$\begin{aligned}
&\arg \min_{\{\mathbf{Q}_d,\mathbf{L},\mathbf{Q}_l,\mathbf{U},\mathbf{Q}_r,\mathbf{R}_s,\mathbf{V}\}} \\
&\|\mathbf{X}_s - \mathbf{Q}_d\mathbf{L}\|_F^2 + \alpha\,\|\mathbf{L} - \mathbf{Q}_l\mathbf{D}_s\|_F^2 + \beta\,\|\mathbf{H} - \mathbf{U}\mathbf{L}\|_F^2 \\
&+ \|\mathbf{X}_s - \mathbf{Q}_d\mathbf{L} - \mathbf{Q}_r\mathbf{R}_s\|_F^2 + \delta\,\|\mathbf{H} - \mathbf{U}\mathbf{L} - \mathbf{V}\mathbf{R}_s\|_F^2 \\
&- \eta\,\|\mathbf{R}_s - \mathbf{W}\mathbf{L}\|_F^2\,, \\
&s.t.\ \mathbf{W} = \arg\min_{\mathbf{W}} \|\mathbf{R}_s - \mathbf{W}\mathbf{L}\|_2^2,\ \|\mathbf{w}_i\|_2^2 \leq 1,\ \|\mathbf{q}_{di}\|_2^2 \leq 1, \\
&\|\mathbf{q}_{ri}\|_2^2 \leq 1, \|\mathbf{q}_{li}\|_2^2 \leq 1,\ \|\mathbf{u}_i\|_2^2 \leq 1,\ \|\mathbf{v}_i\|_2^2 \leq 1,\ \forall i.
\end{aligned} \tag{8}$$

In the above formulation, the second, the third, and the fourth lines are corresponding to the first, the second, and the third criteria, respectively. As the proposed classifier f makes predictions based on only the defined attributes, the first criterion protects f from being distracted from its classification task. However, defined attributes are usually not equally valuable for classification and some of them are highly correlated. Instead of adopting the defined attributes directly, we employ discriminative latent attributes proposed in [10] for zero-shot classification. $\mathbf{L}$ is latent attributes which are derived from the defined attributes and $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, ...]$ where $\mathbf{h}_i = [0, ..., 0, 1, 0, ..., 0]^T$ is a one hot vector which gives the label of sample i. Thus $\mathbf{U}$ can be regarded as the seen-class classifier in the latent attribute space. For the second criterion, we assume the learned residual attributes suffer little of the above problem and adopt them and the discriminative latent attributes jointly for the classification task. For the third criterion, as we expect the residual attributes discover non-redundant properties, the defined attributes should not be predictive for the residual attributes. $\mathbf{w}_i$ is the i-th column of $\mathbf{W}$.

Optimizing the three criteria simultaneously is challenging as there are several hyper-parameters which are set via cross-validation. Furthermore, it may degrade the performance of f, as f makes predictions based on the defined attributes solely. We divide the optimization problem in Eq. 8 into two subproblems which are optimized separately. In the first subproblem, only the first criterion is considered and we optimize $\mathbf{Q}_d, \mathbf{L}, \mathbf{Q}_l$ and $\mathbf{U}$ to strive for f with higher performance. In the second subproblem, $\mathbf{Q}_d, \mathbf{L}, \mathbf{Q}_l$ and $\mathbf{U}$ are fixed and we optimize $\mathbf{Q}_r, \mathbf{R}_s$ and $\mathbf{V}$ with taking the second and the third criteria into consideration. With our proposed optimization procedure, the cross validation work for hyper-parameters $\{\alpha, \beta, \delta, \eta\}$ is significantly reduced as $\{\alpha, \beta\}$ and $\{\delta, \eta\}$ are cross validated separately.

The First Subproblem. Taking only the first criterion into consideration, Eq. 8 is simplified to be

$$\begin{aligned}
&\arg \min_{\{\mathbf{Q}_d,\mathbf{L},\mathbf{Q}_l,\mathbf{U}\}} \|\mathbf{X}_s - \mathbf{Q}_d\mathbf{L}\|_F^2 + \alpha\,\|\mathbf{L} - \mathbf{Q}_l\mathbf{D}_s\|_F^2 + \beta\,\|\mathbf{H} - \mathbf{U}\mathbf{L}\|_F^2\,, \\
&s.t.\ \ \|\mathbf{q}_{di}\|_2^2 \leq 1, \|\mathbf{q}_{li}\|_2^2 \leq 1,\ \|\mathbf{u}_i\|_2^2 \leq 1,\ \forall i.
\end{aligned} \tag{9}$$

This is the problem proposed in [10]. Equation 9 is not convex for $\mathbf{Q}_d, \mathbf{L}, \mathbf{Q}_l$ and $\mathbf{U}$ simultaneously, but it is convex for each of them separately. An alternating optimization method is adopted to solve it. Detailed optimization process can be found in [10].

The Second Subproblem. After solving the first subproblem, $\mathbf{Q}_d, \mathbf{L}, \mathbf{Q}_l$ and $\mathbf{U}$ are fixed and Eq. 8 is simplified to be

$$\arg \min_{\{\mathbf{Q}_r, \mathbf{R}_s, \mathbf{V}\}} \|\mathbf{X}_s - \mathbf{Q}_d\mathbf{L} - \mathbf{Q}_r\mathbf{R}_s\|_F^2 + \delta \|\mathbf{H} - \mathbf{U}\mathbf{L} - \mathbf{V}\mathbf{R}_s\|_F^2 - \eta \|\mathbf{R}_s - \mathbf{W}\mathbf{L}\|_F^2,$$
$$s.t.\ \mathbf{W} = \arg\min_{\mathbf{W}} \|\mathbf{R}_s - \mathbf{W}\mathbf{L}\|_2^2,\ \|\mathbf{w}_i\|_2^2 \leq 1, \|\mathbf{q}_{ri}\|_2^2 \leq 1,\ \|\mathbf{v}_i\|_2^2 \leq 1,\ \forall i. \tag{10}$$

Similarly, $\mathbf{Q}_r, \mathbf{R}_s$ and $\mathbf{V}$ are optimized by the alternate optimization method. The optimization process is briefly described as follows.

(1) Fix $\mathbf{Q}_r, \mathbf{V}$ and update $\mathbf{R}_s$:

$$\arg\min_{\mathbf{R}_s} \left\|\tilde{\mathbf{X}} - \tilde{\mathbf{Q}}\mathbf{R}_s\right\|_F^2, \tag{11}$$

where

$$\tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X}_s - \mathbf{Q}_d\mathbf{L} \\ \delta(\mathbf{H} - \mathbf{U}\mathbf{L}) \\ -\eta(\mathbf{W}\mathbf{L}) \end{bmatrix},\ \tilde{\mathbf{Q}} = \begin{bmatrix} \mathbf{Q}_r \\ \delta\mathbf{V} \\ -\eta\mathbf{I} \end{bmatrix},$$

and $\mathbf{I}$ is the identity matrix. $\mathbf{R}_s$ has the closed-form solution as

$$\mathbf{R}_s = (\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}})^{-1}\tilde{\mathbf{Q}}^T\tilde{\mathbf{X}}. \tag{12}$$

(2) Fix $\mathbf{R}_s, \mathbf{V}$ and update $\mathbf{Q}_r$:

$$\arg\min_{\mathbf{Q}_r} \|\mathbf{X}_s - \mathbf{Q}_d\mathbf{L} - \mathbf{Q}_r\mathbf{R}_s\|_F^2,\ s.t.\ \|\mathbf{q}_{ri}\|_2^2 \leq 1,\ \forall i. \tag{13}$$

The above problem can be solved by the Lagrange dual and the analytical solution is

$$\mathbf{Q}_r = (\mathbf{X}_s - \mathbf{Q}_d\mathbf{L})\mathbf{R}_s{}^T(\mathbf{R}_s\mathbf{R}_s{}^T + \boldsymbol{\Lambda})^{-1}, \tag{14}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix constructed by all the Lagrange dual variables.

(3) Fix $\mathbf{R}_s, \mathbf{Q}_r$ and update $\mathbf{V}$:

$$\arg\min_{\mathbf{V}} \|\mathbf{H} - \mathbf{U}\mathbf{L} - \mathbf{V}\mathbf{R}_s\|_F^2,\ s.t.\ \|\mathbf{v}_i\|_2^2 \leq 1,\ \forall i. \tag{15}$$

The above problem can be solved in the same way as Eq. 13 and the solution is

$$\mathbf{V} = (\mathbf{H} - \mathbf{U}\mathbf{L})\mathbf{R}_s{}^T(\mathbf{R}_s\mathbf{R}_s{}^T + \boldsymbol{\Lambda})^{-1}. \tag{16}$$

(4) Computing $\mathbf{W}$:

$$\arg\min_{\mathbf{W}} \|\mathbf{R}_s - \mathbf{W}\mathbf{L}\|_F^2,\ s.t.\ \|\mathbf{w}_i\|_2^2 \leq 1,\ \forall i. \tag{17}$$

Similar to Eqs. 14 and 16, we can get the solution

$$\mathbf{W} = \mathbf{R}_s\mathbf{L}^T(\mathbf{L}\mathbf{L}^T + \boldsymbol{\Lambda})^{-1}. \tag{18}$$

The complete algorithm is summarized in Algorithm 1. The optimization process usually converges quickly, after tens of iterations in our experiments.

Algorithm 1. Augmented Attribute Learning for Selective ZSC

```
Input: $\mathbf{X}_s, \mathbf{D}_s, \mathbf{H}, \alpha, \beta, \delta, \eta, K_r$
Output: $\mathbf{Q}_d, \mathbf{L}, \mathbf{Q}_l, \mathbf{U}, \mathbf{Q}_r, \mathbf{R}_s, \mathbf{V}$
1: Optimizing the first subproblem according to [10], obtaining $\mathbf{Q}_d, \mathbf{L}, \mathbf{Q}_l, \mathbf{U}$.
2: Fixing $\mathbf{Q}_d, \mathbf{L}, \mathbf{Q}_l, \mathbf{U}$ obtained from 1, and initializing $\mathbf{Q}_r, \mathbf{V}, \mathbf{W}$ randomly according to $K_r$.
3: while not converge do
4:     Optimizing $\mathbf{R}_s$ according to Eq. 12.
5:     Optimizing $\mathbf{Q}_r$ according to Eq. 14.
6:     Optimizing $\mathbf{V}$ according to Eq. 16.
7:     Optimizing $\mathbf{W}$ according to Eq. 18.
8: return $\mathbf{Q}_d, \mathbf{L}, \mathbf{Q}_l, \mathbf{U}, \mathbf{Q}_r, \mathbf{R}_s, \mathbf{V}$
```

6 Experiments

6.1 Datasets and Settings

Datasets. We conduct experiments on three benchmark image datasets for ZSC, including aPascal&aYahoo (**aP&Y**) [29], Animals with Attributes (**AwA**) [1] and Caltech-UCSD Birds-200-2011 (**CUB-200**) [30]. For all the datasets, we split the categories into seen and unseen sets in the same way as [10]: (1) There are two attribute datasets in **aP&Y**: aPascal and aYahoo. These two datasets contains images from disjoint object classes. The categories in aPascal dataset are used as seen classes and those in aYahoo as the unseen ones. (2) **AwA** contains 50 categories, 40 of which are used as seen categories, and the rest 10 are used as the unseen ones. (3) **CUB-200** is a bird dataset for fine-grained recognition. It contains 200 categories, of which 150 are used as seen categories and the rest 50 as the unseen ones. For all the datasets, we adopt the pre-trained VGG19 [31] to extract features.

Cross Validation. There are several hyper-parameters (including γ, $K_r, \alpha, \beta, \delta, \eta$) which are set via cross-validation. As aforementioned, our proposed optimization procedure relaxes the laborious cross-validation work by decomposing the original problem into two subproblems. α, β are firstly optimized on the validation data independent of the others. After that, to further relax the cross-validation work, we optimize δ, η, K_r independent of γ. Finally, γ is optimized. In this paper, we adopt five-fold cross-validation [17] for all these parameters.

Evaluation Metrics. The performance of the classifier is quantified using *coverage* and *risk*. The coverage is defined to be the probability mass of the non-rejected region in $\mathcal{X}_u$ (the feature space of unseen classes)

$$coverage(f, g) \triangleq E_p[g(\mathbf{x})], \tag{19}$$

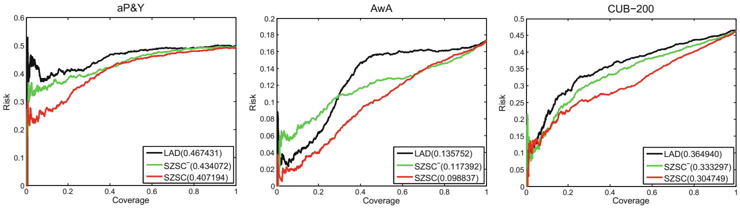

Fig. 2. Comparisons among different variants of the proposed method (best viewed in color). AURCC is given in brackets.

and the selective risk of (f, g) is

$$risk(f, g) \triangleq \frac{E_p[\ell(f(\mathbf{x}), y)g(\mathbf{x})]}{coverage(f, g)}, \tag{20}$$

where ℓ is defined to be 0/1 loss. The risk can be traded off for coverage. Thus the overall performance of a selective classifier can be measured by its Risk-Coverage Curve (RCC), where risk is defined to be a function of the coverage [8,28]. The Area Under Risk-Coverage Curve (AURCC) is usually adopted to quantify the performance.

6.2 Ablation Study

The Effectiveness of Three Criteria. We have incorporated three criteria into the learning of augmented attributes. In this section, we validate the effectiveness of them. We make comparisons among three variants of the proposed method. For the first one, only the first criterion is considered. In other words, no residual attributes are learned, and the classification model degrades to LAD [10] ($conf = conf_d$). For the second one (dubbed SZSC$^-$), the first and the second criteria are considered. For the third one (dubbed SZSC), all the three criteria are incorporated. For all the three variants, the dimensions of the residual attributes are kept the same as that of the defined attributes ($K_r = K_d$). Other hyper-parameters are set via cross-validation. The risk-coverage curves on all the three benchmark datasets are depicted in Fig. 2. It can be seen that on all the three datasets, the proposed method achieves the best performance when all the three criteria are involved.

Trade-Off Between Two Confidence Scores. The proposed confidence function is composed of two parts: the confidence defined within the defined attributes ($conf_d$) and the confidence defined with the aid of the residual attributes ($conf_r$). In this section, we test how the trade-off parameter λ affects the performance of SZSC. If $\lambda = 0$, the confidence depends entirely on the defined attributes. On the contrary, if $\lambda = 1$, the confidence is composed of $conf_r$ only. All other hyper-parameters are kept the same for fair comparisons.

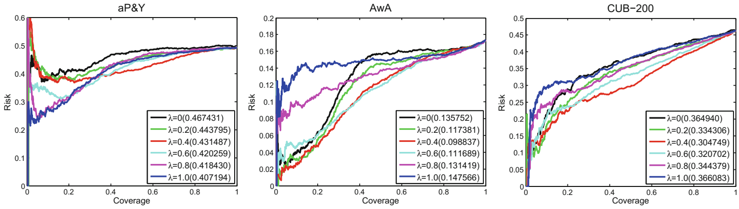

Fig. 3. Risk-coverage curves of the proposed method with varying λ.

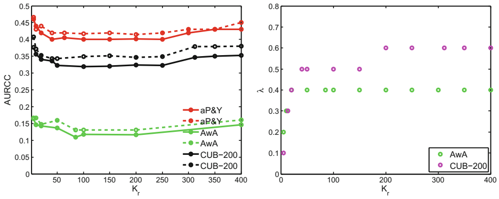

Fig. 4. How AURCC (left) and the optimal λ (right) change with varying K_r.

Experimental results on all the three benchmark datasets are shown in Fig. 3. It reveals that the appropriately combined confidence significantly improve the classifier's performance on all the three datasets. More surprisingly, on aP&Y the optimally combined confidence relies heavily on $conf_r$ ($\lambda = 1.0$). The under-complete defined attribute vocabulary and the large difference between the seen and the unseen categories may account for that.

Dimensions of the Residual Attribute Space. In this section, we investigate how the performance changes with the varying K_r. Similarly, all other hyper-parameters are kept the same. For a more comprehensive view of the proposed method, both SZSC$^-$ and SZSC are evaluated. Experimental results are shown in Fig. 4 (left). It can be observed that the number of dimensions of the residual attribute space also makes unneglected impacts on the final performance. Too small K_r (< 50) will leave the residual discriminative properties not fully explored. Conversely, too large K_r (> 300) renders the optimization more challenging and time-consuming. Both these two cases degrade the performance. Furthermore, we test that how the cross-validated λ changes with K_r. Results are depicted in Fig. 4 (right). It can be seen that with small K_r, the confidence obtained via the residual attributes is unreliable and the optimally combined confidence relies heavily (small λ) on the $conf_d$. However, as K_r becomes larger, λ also becomes larger which indicates that $conf_r$ plays a more important role.

6.3 Benchmark Comparison

Competitors. Several existing ZSC models are selected for benchmark comparison, including SSE [17], SYNC [19], SCoRe [32], SAE [33] and LAD [10].

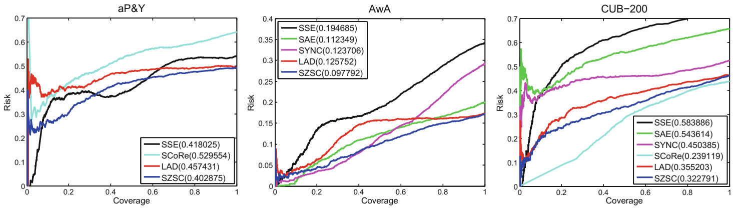

Fig. 5. Risk-coverage curves of existing methods and SZSC.

The selection criteria are (1) representativeness: they cover a wide range of models; (2) competitiveness: they clearly represent the state-of-the-art; (3) recent work: all of them are published in the past three years; (4) reproducibility: all of them are code available, so the provided results in this paper are reproducible. We briefly review them and introduce their typical confidence functions as follows. SSE adopts SVM as the classification model. The margin in SVM classifier is employed as the confidence. SCoRe utilizes deep neural networks integrated with a softmax classifier for ZSC. The softmax output is usually employed for misclassification or out-of-distribution example dection [34]. We also use it as the proxy of the confidence for SCoRe. For the other competitors, the classification task is usually accomplished via nearest neighbor searches in the shared embedding space. We take the cosine similarity as the confidence.

For fair comparisons, both the proposed method and the competitors are tested with features extracted by VGG19. Experimental results are provided in Fig. 5. From the figure, we can conclude that: (1) Many existing ZSC methods exhibit poor performance in Selective ZSC settings. With lower coverage, these classifiers are expected to yield higher accuracy (i.e., lower risk). However, many methods violate that regularity in many cases, especially on aP&Y and CUB. These experimental results give us a more comprehensive view of existing ZSC methods. (2) The proposed method outperforms most existing methods significantly on all the three benchmark datasets. One exception is that SCoRe which utilizes deep neural networks behaves better on CUB-200. However, it produces a much worse performance on aP&Y, as there is a large imbalance among the number of images in different categories (51–5071). (3) Although bringing some improvement, the proposed method remains far behind the ideal. It indicates that there still exists large space for further study.

Augmenting the Self-awareness of Existing Methods. The proposed method focuses on augmenting the defined attributes with residual properties to improve zero-shot performance in selective classification settings. It is orthogonal to how to exploit the defined attributes for ZSC. Thus the proposed method can be combined with most existing attribute-based methods to improve their performance in Selective ZSC settings. Here we propose a simple combining strategy: the confidence functions of existing ZSC methods are directly combined with the proposed confidence function defined with the aid of residual attributes.

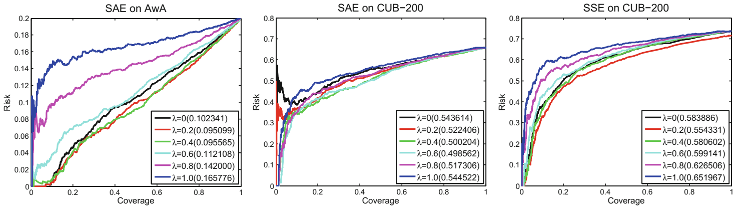

Fig. 6. Combining SZSC with SAE and SSE.

In other words, $conf_r$ is agnostic about the classification model which is used for recognition, and $conf_d$ in Eq. 7 is replaced with the confidence of existing ZSC methods. Experiments are conducted with SAE on AwA and CUB-200 and SSE on CUB-200. Results are shown in Fig. 6. We can see that with the simple proposed combining strategy, the performance of SAE and SSE can be further improved to some degree. These compelling results suggest that the confidence defined by the consistency between the defined and the residual attributes has some generalization ability across ZSC models. We believe learning the residual attributes adaptively with the specified ZSC model (*e.g.*, SCoRe) will further improve the performance, which is left for future research.

7 Conclusions and Future Work

In this paper, we introduce an important yet under-studied problem: zero-shot classifiers can abstain from prediction when in doubt. We empirically demonstrate that existing zero-shot classifiers behave poorly in this new settings, and propose a novel selective classifier to make safer predictions. The proposed classifier explores and exploits the residual properties beyond the defined attributes for defining confidence functions. Experiments show that the proposed classifier achieves significantly superior performance in selective classification settings. Furthermore, it is also shown that the proposed confidence can also augment existing ZSC methods for safer classification.

There are several research lines which are worthy of further study following our work. For example, we propose to learn residual attributes to improve the performance of attribute-based classifiers. Similar ideas may also work for zero-shot classifiers built on word vectors or text descriptions. Another example is that in this paper we propose a straightforward combing strategy to improve the performance of existing methods. We believe learning the residual attributes adaptively with the ZSC model can further improve the final performance. Finally, considering the importance of the proposed selective zero-shot classification problem, we encourage researchers to pay more attention to this new challenge.

Acknowledgements. This work is supported by National Key Research and Development Program (2016YFB1200203), National Natural Science Foundation of China (61572428, U1509206), Fundamental Research Funds for the Central Universities (2017FZA5014) and Key Research, Development Program of Zhejiang Province (2018C01004) and ARC FL-170100117, DP-180103424 of Australia.

References

1. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 951–958. IEEE (2009)
2. Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C.: Label-embedding for attribute-based classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 819–826 (2013)
3. Romera-Paredes, B., Torr, P.: An embarrassingly simple approach to zero-shot learning. In: International Conference on Machine Learning, pp. 2152–2161 (2015)
4. Akata, Z., Reed, S., Walter, D., Lee, H., Schiele, B.: Evaluation of output embeddings for fine-grained image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2927–2936 (2015)
5. Xian, Y., Schiele, B., Akata, Z.: Zero-shot learning - the good, the bad and the ugly. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017
6. Song, J., Shen, C., Yang, Y., Liu, Y., Song, M.: Transductive unbiased embedding for zero-shot learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018
7. Chow, C.K.: An optimum character recognition system using decision functions. IRE Trans. Electron. Comput. EC **6**(4), 247–254 (1957)
8. El-Yaniv, R., Wiener, Y.: On the foundations of noise-free selective classification. J. Mach. Learn. Res. **11**, 1605–1641 (2010)
9. Peng, P., Tian, Y., Xiang, T., Wang, Y., Huang, T.: Joint learning of semantic and latent attributes. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 336–353. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_21
10. Jiang, H., Wang, R., Shan, S., Yang, Y., Chen, X.: Learning discriminative latent attributes for zero-shot classification. In: The IEEE International Conference on Computer Vision (ICCV), October 2017
11. Kan, M., Shan, S., Chen, X.: Multi-view deep network for cross-view classification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016
12. Fu, Y., Hospedales, T.M., Xiang, T., Fu, Z., Gong, S.: Transductive multi-view embedding for zero-shot recognition and annotation. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8690, pp. 584–599. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10605-2_38
13. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al.: Devise: a deep visual-semantic embedding model. In: Advances in neural information processing systems, pp. 2121–2129 (2013)
14. Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C.: Label-embedding for image classification. IEEE Trans. Pattern Anal. Mach. Intell. **38**(7), 1425–1438 (2016)

15. Xian, Y., Akata, Z., Sharma, G., Nguyen, Q., Hein, M., Schiele, B.: Latent embeddings for zero-shot classification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016
16. Socher, R., Ganjoo, M., Manning, C.D., Ng, A.: Zero-shot learning through cross-modal transfer. In: Advances in Neural Information Processing Systems, pp. 935–943 (2013)
17. Zhang, Z., Saligrama, V.: Zero-shot learning via semantic similarity embedding. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4166–4174 (2015)
18. Norouzi, M., et al.: Zero-shot learning by convex combination of semantic embeddings. In: Proceedings of ICLR. Citeseer (2014)
19. Changpinyo, S., Chao, W.L., Gong, B., Sha, F.: Synthesized classifiers for zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5327–5336 (2016)
20. Fu, Y., Xiang, T., Jiang, Y.G., Xue, X., Sigal, L., Gong, S.: Recent advances in zero-shot recognition: toward data-efficient understanding of visual content. IEEE Signal Process. Mag. **35**(1), 112–125 (2018)
21. Berg, T.L., Berg, A.C., Shih, J.: Automatic attribute discovery and characterization from noisy web data. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6311, pp. 663–676. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15549-9_48
22. Sharmanska, V., Quadrianto, N., Lampert, C.H.: Augmented attribute representations. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7576, pp. 242–255. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33715-4_18
23. Rastegari, M., Farhadi, A., Forsyth, D.: Attribute discovery via predictable discriminative binary codes. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7577, pp. 876–889. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33783-3_63
24. Yu, F.X., Cao, L., Feris, R.S., Smith, J.R., Chang, S.F.: Designing category-level attributes for discriminative visual recognition. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, June 2013
25. Szegedy, C., et al.: Intriguing properties of neural networks. In: International Conference on Learning Representations (2014)
26. Mandelbaum, A., Weinshall, D.: Distance-based confidence score for neural network classifiers. CoRR abs/1709.09844 (2017)
27. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on Machine Learning (ICML2016) (2016)
28. Geifman, Y., El-Yaniv, R.: Selective classification for deep neural networks. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems 30, pp. 4885–4894. Curran Associates, Inc. (2017)
29. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 1778–1785. IEEE (2009)
30. Welinder, P., et al.: Caltech-UCSD Birds 200. Technical report CNS-TR-2010-001, California Institute of Technology (2010)
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

32. Morgado, P., Vasconcelos, N.: Semantically consistent regularization for zero-shot recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017
33. Kodirov, E., Xiang, T., Gong, S.: Semantic autoencoder for zero-shot learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017
34. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136 (2016)