Dual-View Ranking with Hardness Assessment for Zero-Shot Learning

Yuchen Guo,[†‡] Guiguang Ding,[‡] Jungong Han,[§] Xiaohan Ding,[‡] Sicheng Zhao,[♮]
Zheng Wang,[♭] Chenggang Yan,[♯] Qionghai Dai[†]
[†]Department of Automation, [‡]School of Software, Tsinghua University, Beijing 100084, China
[§]School of Computing and Communications, Lancaster University, Lancaster, LA1 4YW, UK
[♮]Department of EE and CS, UC Berkeley, USA; [♭]Department of CST, USTB, China
[♯]Institute of Information and Control, Hangzhou Dianzi University, China

Abstract

Zero-shot learning (ZSL) is to build recognition models for previously unseen target classes which have no labeled data for training by transferring knowledge from some other related auxiliary source classes with abundant labeled samples to the target ones with class attributes as the bridge. The key is to learn a similarity based ranking function between samples and class labels using the labeled source classes so that the proper (unseen) class label for a test sample can be identified by the function. In order to learn the function, single-view ranking based loss is widely used which aims to rank the true label prior to the other labels for a training sample. However, we argue that the ranking can be performed from the other view, which aims to place the images belonging to a label before the images from the other classes. Motivated by it, we propose a novel **Du**ra**l**-view **R**an**K**ing (DARK) loss for zero-shot learning simultaneously ranking labels for an image by point-to-point metric and ranking images for a label by point-to-set metric, which is capable of better modeling the relationship between images and classes. In addition, we also notice that previous ZSL approaches mostly fail to well exploit the hardness of training samples, either using only very hard ones or using all samples indiscriminately. In this work, we also introduce a sample hardness assessment method to ZSL which assigns different weights to training samples based on their hardness, which leads to a more accurate and robust ZSL model. Experiments on benchmarks demonstrate that DARK outperforms the state-of-the-arts for (generalized) ZSL.

Introduction

The recent years have witnessed the tremendous progress in zero-shot learning (ZSL) and there is an increasing number of new ZSL approaches every year (Xian, Schiele, and Akata 2017). ZSL is an emerging task of object recognition and image classification whose goal is to recognize categories whose visual exemplars are not given for model training (Lampert, Nickisch, and Harmeling 2014). It is a practical learning task in real-world applications, like Web-image classification, because there are potentially infinite number of categories and new concepts may emerge every day. In addition, since the exemplars for different categories follow a long-tailed distribution such that many uncommon categories have limited visual samples (Changpinyo et al. 2016). In these scenarios, it becomes difficult to collect sufficient labeled training samples for these categories to train recognition models in a conventional supervised learning way (Bishop and others 2006) and thus necessitates ZSL.

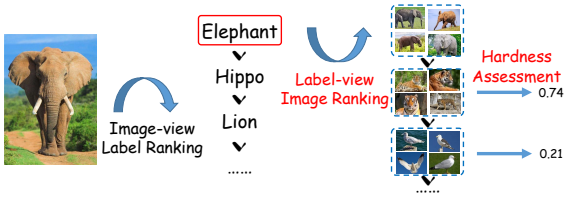


Figure 1: Dual-view ranking with hardness assessment.

The problem of ZSL can be described as: how to tell whether a sample belongs to a previously unseen class? If there is no information about the unseen class, this problem is obviously unsolvable. The key assumption in previous ZSL literatures is that each class label can be represented as a label feature vector, such as manually defined attribute vector (Farhadi et al. 2009; Lampert, Nickisch, and Harmeling 2014), word embedding representation (Socher et al. 2013), or their combinations (Fu et al. 2015b; Akata et al. 2015). Now denote $x \in \mathbb{R}^p$ is the p-dimensional image feature vector like the deep feature (He et al. 2016), and $y_j \in \mathbb{R}^q$ is the q-dimensional feature vector for class c_j. Zero-shot recognition is performed by computing the cross-modality similarity between these features (Xian, Schiele, and Akata 2017):

$$c(x) = \text{argmax}_{c_j} F(\phi(x), \varphi(y_j), \theta_F) \quad (1)$$

where ϕ and φ are the image specific and label specific feature transformations respectively, F is the similarity function and θ_F is the function parameter of F. One simple function is the linear similarity $F(\phi(x), \varphi(y_j), \theta_F) = xWy'_j$. Given some auxiliary related classes with many labeled samples as knowledge source, one can train the model (ϕ, φ, and θ_F) by minimizing a loss function. Since the training source classes and the unseen target classes are related in the label feature space, the function learned with source classes can be transferred and applied to the target classes. Then with the learned function F and the feature vector y of an unseen class, the image-class similarity is produced by F and the classification can be performed, even though there is no labeled visual samples of this class available for training.

Observation and Contribution

The loss function for training F is important for ZSL, which is the focus of many previous works. One widely used loss function which yields state-of-the-art performance is the ranking based loss. The basic idea of ranking based loss is to make the similarity between image x_i and its true class' label feature vector y_i larger than the one between other classes' feature $y \neq y_i$, which can be formulated as below,

$$\min_{\phi,\varphi,\theta_F} \mathcal{L} = \sum_i \sum_{y \neq y_i} [\epsilon(x_i, y_i, y) + F(\phi(x_i), \varphi(y), \theta_F) - F(\phi(x_i), \varphi(y_i), \theta_F)] \cdot \Delta(x_i, y_i, y) + R(F) \tag{2}$$

where $\epsilon(\cdot)$ is a data-dependent margin for ranking, $\Delta(\cdot)$ is a data-dependent weight for the triplet (x_i, y_i, y), and $R(F)$ is a regularization term to reduce model complexity. This general formulation can be specified to many state-of-the-art ZSL approaches, such as DEVISE (Frome et al. 2013), SJE (Akata et al. 2015), SYNC (Changpinyo et al. 2016), ALE (Akata et al. 2016), and LATEM (Xian et al. 2016) if the detailed definitions of ϕ, φ, θ_F, ϵ, Δ, and F are given.

The extraordinary performance and widely usage of ranking based loss for ZSL motivate us to closely investigate it. By summarizing previous works, we can observe that existing approaches mostly fail to consider the following issues:

1. *Dual-view ranking*. Most of existing approaches only take single-view ranking into account. They perform image-view label ranking which aims to rank the true label before any other labels for a training image. However, the other view, i.e., label-view image ranking, is always ignored, which aims to rank the images belonging to a label before images from other classes. Intuitively, zero-shot classification shown in Eq. (1) is a cross-modality matching procedure, where considering the information from both views simultaneously usually leads to better ranking performance (Socher et al. 2014; Karpathy, Joulin, and Li 2014).

2. *Hardness assessment*. In previous works, hard triplet mining is always adopted, like setting $\Delta(x_i, y_i, y) = 1$ if $\epsilon(x_i, y_i, y) + F(\phi(x_i), \varphi(y), \theta_F) - F(\phi(x_i), \varphi(y_i), \theta_F) > 0$ or 0 otherwise. Hard training sample mining has been shown to be an effective way for learning to rank (Schroff, Kalenichenko, and Philbin 2015) because more attention is paid to data on which the current model has large loss. However, it is sometimes sensitive to noise and outliers since they always lead to large error such that the model may bias to them from the true distribution. On the other hand, using all triplets indiscriminately may lead to slow convergence and bad optimum as it suffers from imbalanced easy-hard distribution (Shrivastava, Gupta, and Girshick 2016). Therefore, using hardness assessment instead of either select-or-not or select-all strategies seems more reasonable in practice.

To combine the above issues with ranking based ZSL, in this paper we propose a novel **Du**a**l**-view **R**an**K**ing with hardness assessment (DARK) for ZSL. In particular, we adopt a dual-view ranking loss which simultaneously performs image-view label ranking by point-to-point metric and label-view image ranking by point-to-set metric (Zhou et al. 2017). By dual-view ranking, the similar relationship between images and labels can be better exploited. In addition, we utilize hardness assessment to assign different weights to training triplets based on their errors under the current model. In this way, the information of training data can be fully exploited. In summary, we make the contributions below:

1. Noticing that the majority of existing ranking based ZSL approaches consider image-view label ranking only, we propose a novel dual-view ranking loss which jointly minimizes the image-view label ranking loss and label-view image ranking (DARK) loss, where a point-to-point metric and a point-to-set metric are adopted for them respectively considering the data properties in different views. To our best knowledge, it is the first ZSL work using dual-view ranking.

2. Different from select-or-not or select-all strategies for choosing training triplets, we apply a hardness assessment method to triplets and then assign different training weights to different triplets based on the error under the current model, which better captures the knowledge in training data.

3. We conduct extensive experiments on four benchmark datasets for (generalized) zero-shot classification. Empirical results demonstrate DARK outperforms the state-of-the-arts.

Table 1: Notations and descriptions.

Notation	Description	Notation	Description
x^s, x^u	image feature	n_s, n_u	#sample
c^s, c^u	class label	p	#dimension
y_c	class feature	q	#class feature
W, U, V	matrix	k_s, k_u	#class

Preliminary and Related Work

Preliminary and Notation

The problem of ZSL is defined as follows. we have two disjoint class sets $\mathcal{C}^s = \{c_1^s, ..., c_{k_s}^s\}$ and $\mathcal{C}^u = \{c_1^u, ..., c_{k_u}^u\}$ denoting seen classes and unseen classes respectively with $\mathcal{C}^s \cap \mathcal{C}^u = \emptyset$, where k_s and k_u are the number of classes in seen set and unseen set. From the image perspective, a training image set $\mathcal{D}^s = \{x_1^s, ..., x_{n_s}^s\}$ is given where each image x_i^s is associated with one seen class $c_i^s \in \mathcal{C}^s$. We use $\{x_i^s, y_i^s\}_{i=1}^{n_s}$ to learn the similarity function between images and classes. At the test stage, an image x^u is given and our goal is to predict its class in the unseen classes $c^u \in \mathcal{C}^u$, which is the standard setting of ZSL. In the generalized ZSL (GZSL) (Xian et al. 2017), an image x^t is given which is from $\mathcal{C}^s \cup \mathcal{C}^u$, unlike the standard ZSL. The classification is performed by Eq. (1). To enable similarity measure between images and classes, each class $c \in \mathcal{C}^s \cup \mathcal{C}^u$ has a feature vector $y_c \in \mathbb{R}^q$, which can be class attribute vector (Lampert, Nickisch, and Harmeling 2014) or word2vec output (Socher et al. 2013). We show some important notations in Table 1.

Related Work

As introduced above, many existing ZSL approaches with ranking based loss can be summarized as Eq. (2) with different function settings. For simplicity, denote $\mathcal{R}(x_i, y_i, y) = \epsilon(x_i, y_i, y) + F(\phi(x_i), \varphi(y), \theta_F) - F(\phi(x_i), \varphi(y_i), \theta_F)$. DEVISE (Frome et al. 2013) sets $F(\phi(x), \varphi(y), \theta_F) = xWy'$, $\epsilon(x_i, y_i, y) = 1$ if $y \neq y_i$ or 0 otherwise, and $\Delta(x_i, y_i, y) =$

1 if $\mathcal{R}(x_i, y_i, y) > 0$ or 0 otherwise. LDF (Li et al. 2018) use similar definitions like DEVISE. ALE (Akata et al. 2016) has similar settings like DEVISE but sets $\Delta(x_i, y_i, y)$ to a positive ranking based score instead of 1 for triplets with $\mathcal{R}(x_i, y_i, y) > 0$. For SJE (Akata et al. 2015), $\Delta(x_i, y_i, y) = 1$ if $\mathcal{R}(x_i, y_i, y) > 0 \wedge y = \text{argmax}_y \mathcal{R}(x_i, y_i, y)$ or 0 otherwise. LATEM (Xian et al. 2016) adopts multiple matching matrices W_m and defines $F(\phi(x), \varphi(y), W_m) = \max_{1 \leq m \leq M} \phi(x) W_m \varphi(y)'$. SYNC (Changpinyo et al. 2016) defines $F(\phi(x), \varphi(y), \theta_F) = x\varphi(y)'$ and $\Delta(x_i, y_i, y)$ like SJE, where φ is a label feature transformation based on phantom classes. SSE (Zhang and Saligrama 2015) utilizes sparse coding for ϕ and intersection function or rectified linear unit for φ, and defines $F(\phi(x), \varphi(y), \theta_F) = \phi(x)\varphi(y)'$. Obviously, it can be observed that these approaches only consider the image-view ranking and utilize hard triplet mining with select-or-not strategy. For example, SJE selects only the hardest negative label to construct the triplet for model training, whereas the other triplets have no effect on the loss.

Apart from ranking based loss, there is another line adopting regression based loss which aims to generate large similarity for (x_i, y_i) and small similarity for $(x_i, y \neq y_i)$. Representative works include CMT (Socher et al. 2013) with squared loss between $\phi(x)$ and y, EZZSL (Romera-Paredes and Torr 2015) with squared loss between xWy' and one-hot label vector, SAE (Kodirov, Xiang, and Gong 2017) with squared loss from xW to y and from yW' to x, Deep-SCoRe (Morgado and Vasconcelos 2017) with cross-entropy loss between $\phi(x)Wy'$ and one-hot label vector, GSC-Net-SLE (Wu et al. 2018) with cross entropy loss between $\phi(x)W^s$ and the one hot vector of seen classes and soft cross entropy loss between $\phi(x)W^u$ and any y_j^u, and EXEM (Changpinyo, Chao, and Sha 2017) on the other hand consider the distance between $\varphi(y)W'$ and x_c where x_c is the center of features belonging to class c. Regression based loss mainly focuses on only the absolute value of $F(\phi(x), \varphi(y), \theta_F)$. However, we can notice that the classification step in Eq. (1) is fundamentally a ranking operation in which case the relative similarity score is more important.

There are some complicated ZSL approaches proposed in recent years, like sample transfer based ZSL (Guo et al. 2017b), sample synthesis based ZSL (Jurie, Bucher, and Herbin 2017), GAN based ZSL (Zhu et al. 2018), and etc. It is easy to observe that they also rely on the class-image similarity $F(\phi(x), \varphi(y), \theta_F)$ in an implicit or explicit way.

Dual-view Ranking with Hardness Assessment

Objective Function

DARK also follows Eq. (1) for zero-shot classification and thus its goal is to train a similarity function F. As discussed above, ranking based loss achieves state-of-the-art performance such that we also base DARK on ranking based loss. Different from previous single-view ranking loss, in this work we further consider two extra issues, dual-view ranking which additionally considers the label-view image ranking, and hardness assessment which assigns different training weight to different triplets based on their current loss.

Dual-view Ranking. The similarity function F performs cross-modality matching between image feature and label feature. Intuitively, when training F, simultaneously considering image-view label ranking and label-view image ranking is beneficial for training a good image-label matching model, which inspires us to propose the dual-view ranking. In particular, the image-view label ranking focuses on ranking the true label of an image before any other labels, i.e., $F(\phi(x_i), \varphi(y_i), \theta_F) > F(\phi(x_i), \varphi(y), \theta_F)$ for $\forall y \neq y_i$. To achieve this goal, one can use labeled samples from the seen classes to train F by minimizing the loss function in Eq. (2).

We also follow this general loss function but make two modifications. The first is that we use hardness assessment for $\Delta(x_i, y_i, y)$ which will be introduced later instead of the select-or-not strategy with hard triplet mining. The second is that we use a density adaptive margin ϵ instead of a constant margin used in previous approaches (Frome et al. 2013; Akata et al. 2016; Xian et al. 2016; Li et al. 2018). In particular, we use the $\mathrm{softplus}$ function to define it as follows:

$$\epsilon_{DA}(x_i, y_i, y) = m \cdot \log(1 + exp(F(\phi(x_i), \varphi(y_i), \theta_F))) \tag{3}$$

where $m \in (0, 1)$ is a constant to control the scale. The reason why we use density adaptive margin instead of a constant one is as follows. It has been observed that image data has various density in different parts of feature space (Socher et al. 2013; Guo et al. 2017a). In addition, different images may have totally different similarity scores since they have different properties. It is natural that the same margin for different data may have different meanings. For example, an image might have small similarity scores to all labels because there are just a few similar labels to its true label such that a small margin is sufficient to distinguish between its true label and other labels, while another image might have large similarity scores to many labels since there are many similar labels to it such that the same margin is not discriminative enough. So it is more reasonable if the margin is set based on the data itself. In addition, if we use the approximation $\log(1 + e^x) \approx x$, we have $\mathcal{R}(x_i, y_i, y) \approx F(\phi(x_i), \varphi(y), \theta_F) - (1 - m)F(\phi(x_i), \varphi(y_i), \theta_F)$. To minimize the loss, the model should have larger value for the second term because of the factor $1 - m$. This idea is very similar to the large-margin loss (Liu et al. 2016) which has been demonstrated to result in a more discriminative model.

From the other view, we propose to perform label-view image ranking too. In this view, we hope that the images belonging to a class c have larger score to y_c than the images from the other classes. Moreover, unlike the image-view ranking which uses a point-to-point metric, we propose to use a point-to-set metric (Zhou et al. 2017) instead for label-view ranking. Analogous to Eq. (2), the ranking loss function for label-view image ranking is defined as below,

$$\min_{\phi, \varphi, \theta_F} \mathcal{L}_{label} = \sum_{c \in \mathcal{C}^s} \sum_{\hat{c} \neq c} [\epsilon(y_c, \mathcal{S}_c, \mathcal{S}_{\hat{c}}) + F_{set}(\mathcal{S}_{\hat{c}}, \varphi(y_c), \theta_F) - F_{set}(\mathcal{S}_c, \varphi(y_c), \theta_F))] \cdot \Delta(y_c, \mathcal{S}_c, \mathcal{S}_{\hat{c}}) + R(F) \tag{4}$$

where $\mathcal{S}_c$ denotes all images belonging to label c, and F_{set} is a point-to-set metric between a point y to a set of data $\mathcal{S}$.

Since a label vector represents the general semantic characteristics of a class, it is more reasonable to use all images of this class to compute the point-to-set similarity for ranking where the label is matched to the general visual information of all images. The point-to-set metric is defined as follows:

$$F_{set}(\mathcal{S}_{\hat{c}}, \varphi(y_c), \theta_F) = \sum_{x \in \mathcal{S}_{\hat{c}}} w_x F(\phi(x), \varphi(y_c), \theta_F) \quad (5)$$

where F is the similarity function introduced before, and w_x is the weight for image x. In particular, it is expected the images which are highly representative for a class to have large weight because they can capture the general characteristics of this class, which leads to the following definition for w_x:

$$w_x = \frac{1}{Z} exp\{-||x - \frac{1}{|\mathcal{S}_{\hat{c}}|} \sum_{\hat{x} \in \mathcal{S}_{\hat{c}}} \hat{x}||_2^2\} \quad (6)$$

where $Z = \sum_{\hat{x} \in \mathcal{S}_{\hat{c}}} w_{\hat{x}}$ is a normalization factor. This definition considers the distance between x and the class' visual center and assigns large weight to the images which are close to the center indicating they are representative for this class.

In addition, the margin $\epsilon(y_c, \mathcal{S}_c, \mathcal{S}_{\hat{c}})$ is defined in a density adaptive way like Eq. (3) by simply replacing F with F_{set}.

With the image-view label ranking loss in Eq. (2) denoted as $\mathcal{L}_{image}$ and the label-view image ranking loss in Eq. (4), the proposed dual-view ranking loss is the weighted average of them which considers both parts simultaneously and optimizes them jointly. The definition is given as follows:

$$\mathcal{L}_{dual} = \frac{1}{n_s} \mathcal{L}_{image} + \frac{1}{k_s} \mathcal{L}_{label} \quad (7)$$

Hardness Assessment. The dual-view loss in Eq. (7) is based on triplet loss and we can directly use hard triplet mining like in many previous works. In this work, we propose to perform hardness assessment for triplets which assigns large weight for hard triplets and small weight for easy triplets. The widely used hard triplet mining adopts a select-or-not strategy which ignores easy triplets. However, hard triplet mining focuses mainly on the large-loss triplets which are more likely to be noise and outliers. On the other hand, the general knowledge about a class is shared by the majority of the images belonging to this class. Therefore, if the model yields small loss for most of the images, we can believe that it has captured the important information to recognize this class. In this case, hard triplet mining may push the model away from the best one and to fit the outliers, which is unexpected. On the contrary, if we assign the same weight to all triplets, no matter how hard they are, the easy ones may dominate the loss function such that the training procedure converges slowly. To take the advantages from both sides into account, it is a reasonable choice to perform hardness assessment for triplets and assign different weights to them.

For simplicity, we denote $\mathcal{R}(x_i, y_i, y) = \epsilon(x_i, y_i, y) + F(\phi(x_i), \varphi(y), \theta_F) - F(\phi(x_i), \varphi(y_i), \theta_F)$. Obviously, a larger $\mathcal{R}$ means larger loss, indicating a harder triplet. Then we use the $\mathrm{sigmoid}$ function to define the training weight:

$$\Delta(x_i, y_i, y) = (1 + exp(-\mathcal{R}(x_i, y_i, y)))^{-1} \quad (8)$$

For $\Delta(y_c, \mathcal{S}_c, \mathcal{S}_{\hat{c}})$, we analogously define $\mathcal{R}(y_c, \mathcal{S}_c, \mathcal{S}_{\hat{c}}) = \epsilon(y_c, \mathcal{S}_c, \mathcal{S}_{\hat{c}}) + F_{set}(\mathcal{S}_{\hat{c}}, \varphi(y_c), \theta_F) - F_{set}(\mathcal{S}_c, \varphi(y_c), \theta_F)$.

Algorithm 1 Training DARK

Require: Training data $x_i, y_i \in \mathcal{C}^s$;
Ensure: Model parameters U and V;
1: Randomly initialize U and V;
2: Compute w_x by Eq. (6)
3: **repeat**
4: Compute current embeddings $x_i U$ and $y_c V$;
5: Compute ϵ_{DA}^{image} and ϵ_{DA}^{label} by Eq. (3);
6: Compute Δ_{image} and Δ_{label} by Eq. (8);
7: Update $U = U - \tau \frac{\partial \mathcal{L}_{DARK}}{\partial U}$ with Eq. (10);
8: Update $V = V - \tau \frac{\partial \mathcal{L}_{DARK}}{\partial V}$ with Eq. (11);
9: **until** Convergence or maximum iterations;
10: Return U and V;

The $\mathrm{sigmoid}$ function looks like the step function in hard triplet mining. It approaches to 1 for hard triplets and 0 for easy triplets. However, $\mathrm{sigmoid}$ function is softer around the boundary, i.e, $\mathcal{R} \approx 0$. More importantly, it assigns small weight to easy triplets instead of 0 such that the easy triplets still have influence on the loss function. Due to the large number of easy triplets, their information can be captured by the model based on our method while it is totally ignored if their weight is 0 like in conventional hard triplet mining. In this way, the information in both hard and easy triplets are utilized for training at the same time, while the select-or-not strategy fails to achieve this goal by ignoring the easy ones.

Overall Objective Function. With the dual-view ranking in Eq. (7), the hardness assessment in Eq. (8), and the density-adaptive margin in Eq. (3), we obtain the overall objective function of DARK, which is formulated as below:

$$\begin{aligned}
&\min_{\phi, \varphi, \theta_F} \mathcal{L}_{DARK} \\
&= \frac{1}{n_s} \sum_i \sum_{y \neq y_i} [\epsilon_{DA}^{image}(x_i, y_i, y) + F(\phi(x_i), \varphi(y), \theta_F) \\
&- F(\phi(x_i), \varphi(y_i), \theta_F)] \cdot \Delta_{image}(x_i, y_i, y) \\
&+ \frac{1}{k_s} \sum_{c \in \mathcal{C}^s} \sum_{\hat{c} \neq c} [\epsilon_{DA}^{label}(y_c, \mathcal{S}_c, \mathcal{S}_{\hat{c}}) + F_{set}(\mathcal{S}_{\hat{c}}, \varphi(y_c), \theta_F) \\
&- F_{set}(\mathcal{S}_c, \varphi(y_c), \theta_F))] \cdot \Delta_{label}(y_c, \mathcal{S}_c, \mathcal{S}_{\hat{c}}) + R(F)
\end{aligned} \quad (9)$$

In particular, since we use dual-view ranking instead of single view ranking, we propose to use image specific projection and label projection respectively, i.e, we have $\phi(x) = xU$ and $\varphi(y) = yV$ where $U \in \mathbb{R}^{p \times r}$ is the image-view linear projection and $V \in \mathbb{R}^{q \times r}$ is the label-view linear projection. We define $F(\phi(x), \varphi(y), \theta_F) = \phi(x)\phi(y)'$. One may notice that if we let $W = UV'$, we have $F(\phi(x), \varphi(y), \theta_F) = xWy'$ which seems simpler. In fact, DARK focuses on dual-view ranking which considers image embedding and label embedding jointly. On the other hand, xWy' can be regarded as single image embedding problem (i.e., $(xW)y'$) or single label embedding ($x(yW')'$). In addition, as suggested by Akata et al. (2016), if we have $r < min(p, q)$, this decomposition can reduce the number of variables significantly from pq to $(p+q)r$, which simplifies the optimization and improves the model efficiency. For the regularization term, we set $R(F) = \lambda(||U||_F^2 + ||V||_F^2)$.

Optimization

There are two matrix variables in the objective function. Obviously, optimizing them simultaneously is very difficult since this will lead to a non-convex and non-smooth problem. So we propose to use an iterative algorithm which alternatively optimizes one matrix while fixing the other in a gradient descent way. One may notice that the margin ϵ and the weight Δ is dependent on U and V too, which makes the partial derivatives complicated. Since our algorithm is based on gradient descent method, the parameter change in one iteration is quite small, and thus we fix ϵ and Δ as constants when optimizing U and V and update them afterwards. The partial derivatives of $\mathcal{L}_{DARK}$ with respect to U and V are:

$$\frac{\partial \mathcal{L}_{DARK}}{\partial U} = \frac{1}{n_s}\sum_i \sum_{y\neq y_i} x_i'(y-y_i)V \cdot \Delta_{image}(x_i, y_i, y) + \lambda U + \frac{1}{k_s}\sum_{c\in\mathcal{C}^s}\sum_{\hat{c}\neq c}\Big(\sum_{\hat{x}\in\mathcal{S}_{\hat{c}}} w_{\hat{x}}\hat{x}' - \sum_{x\in\mathcal{S}_c} w_x x'\Big) y_c V \cdot \Delta_{label}(y_c, \mathcal{S}_c, \mathcal{S}_{\hat{c}}) \tag{10}$$

$$\frac{\partial \mathcal{L}_{DARK}}{\partial V} = \frac{1}{n_s}\sum_i \sum_{y\neq y_i} (y-y_i)' x_i U \cdot \Delta_{image}(x_i, y_i, y) + \lambda V + \frac{1}{k_s}\sum_{c\in\mathcal{C}^s}\sum_{\hat{c}\neq c} y_c'\Big(\sum_{\hat{x}\in\mathcal{S}_{\hat{c}}} w_{\hat{x}}\hat{x} - \sum_{x\in\mathcal{S}_c} w_x x\Big) U \cdot \Delta_{label}(y_c, \mathcal{S}_c, \mathcal{S}_{\hat{c}}) \tag{11}$$

We use the gradients above to gradually and iteratively update U and V respectively with a tiny step size, e.g., 10^{-3}. We summarize the optimization algorithm in Algorithm 1.

Discussion

Complexity. The time complexity of Algorithm 1 is as follows. The complexity for random initialization (line 1) is $\mathcal{O}(pr + qr)$, for computing w_x (line 2) is $\mathcal{O}(n_s p)$, for computing current embeddings (line 4) is $\mathcal{O}(n_s pr + k_s qr)$, for computing $F(x, y, U, V)$ for all image-class pairs is $\mathcal{O}(n_s k_s r)$ since the embeddings xU and yV are given, for computing $F_{set}(\mathcal{S}_{\hat{c}}, \varphi(y_c), U, V)$ for all $(\hat{c}, c)$ pairs is $\mathcal{O}(n_s k_s)$ given $F(x, y, U, V)$, for computing density adaptive margin (line 5) ϵ_{DA}^{image} and ϵ_{DA}^{label} (line 6) is $\mathcal{O}(1)$ for one image-view triplet or label-view triplet, for computing hardness assessment based weight Δ_{image} and Δ_{label} is also $\mathcal{O}(1)$ for either kind of triplet, for updating U (line 7) by Eq. (10) is $\mathcal{O}(n_s k_s pr)$, and for updating V (line 8) by Eq. (11) is $\mathcal{O}(n_s k_s qr)$. Suppose the objective function converges in T iterations, the total time complexity is $\mathcal{O}(T(n_s k_s(pr + qr + r + 1) + n_s pr + k_s qr) + pr + qr + n_s p)$, which is linear to the number of training samples. In addition, we can observe that the gradients in Eq. (10) and Eq. (11) are image-wise decoupled if $\sum_x \in \mathcal{S}_c w_x x$ is pre-computed, which indicates that the total gradients are the summation of the gradient from each individual image. In this case, we do not need all n_s images in an iteration. Alternatively, we can use mini-batch based optimization which samples a subset of training images (e.g., 256) in one iteration for training. Moreover, since ϵ and Δ do not change dramatically in one iteration considering U and V are updated with tiny stepsize, we do not need to update ϵ and Δ in every iteration. In experiments, we update them every 10 iterations. By using mini-batch gradient descent with b images in every iteration, the complexity is $\mathcal{O}(Tbk_s(pr + qr + r) + bpr + k_s pr)$ for the iterative part (line 3 to 9), which is much more efficient in practice.

Table 2: The statistics of datasets.

	AwA2	aPY	SUN	CUB
#seen class	40	20	645	150
#(train) seen sample	23,527	5,932	10,320	7,057
#(test) seen sample	5,882	1,483	2,580	1,764
#unseen class	10	12	72	50
#unseen sample	7,913	7,924	1,440	2,967
#class attribute	85	64	102	312

Relationship to existing works. The objective function in Eq. (9) is formulated in a general way. In fact, many existing ZSL approaches can be regarded as a special case of DARK. In particular, DARK considers dual-view ranking where the image-view ranking has the same framework as existing works, as summarized in Eq. (2). By setting the weight of $\mathcal{L}_{label}$ to 0, ϵ to a constant like 1, and Δ to step function, Eq. (9) is similar to the objective function of DEVISE (Frome et al. 2013), ALE (Akata et al. 2016), and many other works (Akata et al. 2015; Xian et al. 2016; Li et al. 2018). The difference between DARK and them is that we introduce label-view ranking with point-to-set metric into the loss function. To our best knowledge, this is the first ZSL work that utilizes dual-view ranking. In addition, DARK takes the hardness of triplets into account and adopts soft weight, while previous works mostly adopt select-or-not strategy which ignores the information in many easy triplets.

Experiment

Setting

In the experiment, we utilize four widely used standard benchmark datasets for ZSL. The first dataset is Animals with Attributes2 (AwA2) (Xian et al. 2017) with 50 animal species of which 40 are used as seen classes and the other 10 as unseen classes. The second dataset is aPascal-aYahoo (aPY) (Farhadi et al. 2009). It has 20 classes from Pascal VOC challenge like "person" and "dog" as the seen classes, and 12 related classes like "centaur"" and "wolf" collected from Yahoo search engine. The third dataset is SUN (Patterson and Hays 2012) scene recognition dataset with 717 different scenes of which 645 are used as seen classes and the other 72 as unseen classes. The last dataset is CUB (Wah et al. 2011) bird fine-grained recognition dataset with 200 kinds of birds of which 150 are used as seen classes and the other 50 as unseen classes. For each dataset, some seen class images are used for model training and the other seen class images together with all unseen class images are utilized as the test set. For each image, the ResNet-101 (He et al. 2016) pre-trained on ImageNet is employed as feature extractor producing 2,048-dimensional image feature. For each class, the class attribute vector is regarded as the label feature vector. For fair comparison, we use the seen-unseen split, train-test split, image feature, and label feature given by Xian et al. (2017). The statistics are shown in Table 2.

Table 3: (Generalized) ZSL performance comparison on benchmarks. ZSL is evaluated by ACC and GZSL is evaluated by H.

	AwA2		aPY		SUN		CUB		Average	
	ACC	H	ACC	H	ACC	H	ACC	H	ACC	H
DEVISE (Frome et al. 2013)	59.7	27.8	39.8	9.2	56.5	20.9	52.0	32.8	52.00	22.68
CONSE (Norouzi et al. 2013)	44.5	1.0	26.9	0.0	38.8	11.6	34.3	3.1	36.13	3.93
CMT (Socher et al. 2013)	37.9	15.9	28.0	19.0	39.9	13.3	34.6	8.7	35.10	14.23
SJE (Akata et al. 2015)	61.9	14.4	32.9	6.9	53.7	19.8	53.9	33.6	50.60	18.68
EZZSL (Romera-Paredes and Torr 2015)	58.6	11.0	38.3	4.6	54.5	15.8	53.9	21.0	51.33	13.10
SSE (Zhang and Saligrama 2015)	61.0	14.8	34.0	0.4	51.5	4.0	43.9	14.4	47.60	8.40
ALE (Akata et al. 2016)	62.5	23.9	39.7	8.7	58.1	26.3	54.9	34.4	53.80	23.33
SYNC (Changpinyo et al. 2016)	46.6	18.0	23.9	13.3	56.3	13.4	55.6	19.8	45.60	16.13
LATEM (Xian et al. 2016)	55.8	20.0	35.2	0.2	55.3	19.5	49.3	24.0	48.90	15.93
SAE (Kodirov, Xiang, and Gong 2017)	54.1	2.2	8.3	0.9	40.3	11.8	33.3	13.6	34.00	7.13
PSR (Annadani and Biswas 2018)	63.8	32.3	38.4	21.4	61.4	26.7	56.0	33.9	54.90	28.58
ICINESS (Guo et al. 2018)	64.2	36.3	42.4	23.1	62.9	30.3	59.8	39.4	57.33	32.28
ZKL (Zhang and Koniusz 2018)	**70.5**	30.8	45.3	20.5	61.7	25.1	57.1	35.1	58.65	27.88
DARK	**68.9**	**38.3**	**47.1**	**27.0**	**66.0**	**32.8**	**62.5**	**41.6**	**61.13**	**34.93**

Two ZSL tasks are considered in the experiment. The first task is stardard ZSL, where we only use images from $\mathcal{C}^u$ as test data and the goal is to assign a label $c \in \mathcal{C}^u$ to the test sample. As introduced above, DARK utilizes Eq. (1) for classification. We use the average per-class top-1 accuracy (Xian et al. 2017) to evaluate the performance as follows:

$$ACC_{\mathcal{C}} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\frac{\#\text{correct predictions in } c}{\#\text{samples in } c} \tag{12}$$

where $\mathcal{C} = \mathcal{C}^u$. We use all samples from the seen classes, including both train and test in Table 2, to train the model.

The second task is generalized ZSL (GZSL) where a test sample may come from both $\mathcal{C}^u$ and $\mathcal{C}^s$ and the goal is to assign a label $c \in \mathcal{C}^s \cup \mathcal{C}^u$ to the test sample. Because the model is likely to assign larger values to seen classes (Chao et al. 2016), we slightly modify the prediction as follows:

$$\hat{c}_g(x) = \text{argmax}_{c\in\mathcal{C}^s\cup\mathcal{C}^u} F(\phi(x), \varphi(c), \theta_F) - \gamma\mathbb{I}(c \in \mathcal{C}^s) \tag{13}$$

where we simply use $\gamma = 0.2$ in this paper. In GZSL, the test data contains two parts. One is the unseen samples and the other is the test seen samples (the fourth row in Table 2). We only use the train seen samples (the third row in Table 2) as the training set. The harmonic mean of the accuracies on seen classes and unseen classes is used as evaluation metric:

$$H = \frac{2 \times ACC_{\mathcal{C}^s} \times ACC_{\mathcal{C}^u}}{ACC_{\mathcal{C}^s} + ACC_{\mathcal{C}^u}} \tag{14}$$

Implementation

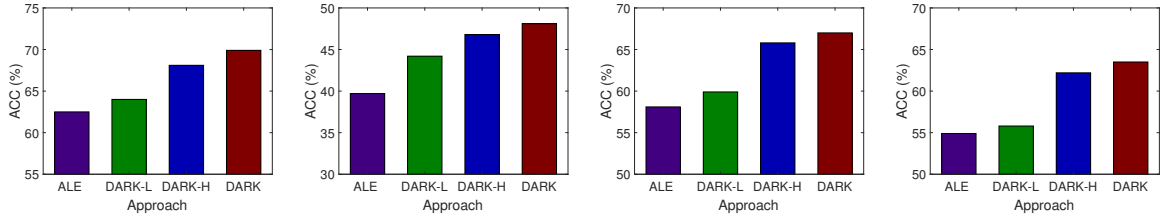
Since we use a linear similarity for F, the scales of x and y have influence on the output score. To remove this influence, we normalize x and y to unit length. When computing density adaptive margin by Eq. (3), we set $m = 0.5$. For the regularization term, we set $\lambda = 0.01$. For mini-batch based gradient descent, each batch contains $b = 512$ images. The margin ϵ and weight Δ are updated every 10 iterations since they change a little in one iteration. The parameter r for U and V are set to 64 consistently. The iteration in Algorithm 1 is conducted for 200 times. The learning rate (stepsize) τ is set to 0.01 initially and then to 0.001 at the 150-th iteration.

Benchmark Comparison

We compare DARK to many related state-of-the-art ZSL approaches on four benchmark for ZSL and GZSL approaches. Since our work focuses on inductive ZSL, some transductive ZSL approaches (Fu et al. 2015a; Kodirov et al. 2015; Guo et al. 2016) are not used as baseline approaches. We use Accuracy (ACC) for standard ZSL performance evaluation and harmonic mean (H) for GZSL performance evaluation.

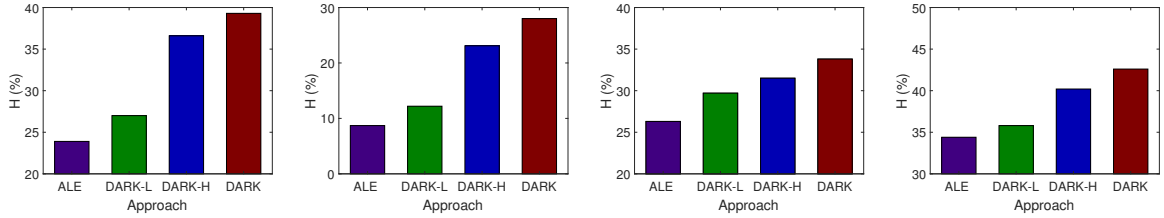
The performance comparison on for benchmarks and two tasks is summarized in Table 3. We can notice that DARK achieves observable improvement over the other baselines in almost all cases. In particular, DARK improves ZSL ACC by 2.48 and GZSL H by 2.65 respectively on average compared to the best baseline approachesThe other baseline approaches mainly adopt image-view label ranking while ignoring the label-view image ranking, and employ hard triplet mining which ignores many relatively easy triplets. To address these issues, we propose to perform dual-view ranking which takes images and labels into account simultaneously. In addition, hardness assessment is applied to all triplets to assign different training weights to them based on their current error such that the information is fully utilized.

We notice that some baseline approaches have some similar ideas to DARK. The first approach is ALE (Akata et al. 2016) which assigns different training weights to triplets based on the current ranking order of labels. The difference between ALE and DARK is that 1) ALE does not consider the label-view ranking and 2) ALE only considers the hard triplets instead of all triplets such that some information is unavoidably lost. Another two approaches are SAE (Kodirov, Xiang, and Gong 2017) and ZKL (Zhang and Koniusz 2018) which consider dual-view information. The difference is below. 1) SAE and ZKL utilize regression loss making x_iW close to y_i and y_iW' close to x_i. However, as shown in Eq. (1), ZSL classification is a ranking problem and it fails to model the ranking information in an explicit way. 2) They treat all training samples with the same weight indiscriminately such that they may be dominated by easy examples.



(a) AwA2 (b) aPY (c) SUN (d) CUB

Figure 2: The effect of dual-view ranking and hardness assessment on ZSL accuracy (ACC).



(a) AwA2 (b) aPY (c) SUN (d) CUB

Figure 3: The effect of dual-view ranking and hardness assessment on GZSL harmonic mean (H).

Ablation Study

It is necessary to verify the effect of dual-view ranking and hardness assessment, which are the contributions of this work, on the performance. In particular, we denote DARK-L[1] as the version which removes the label-view ranking and DARK-H as the version which removes the hardness assessment by setting Δ to step function instead of the sigmoid. If we remove both parts, DARK has only image-view ranking such that it is very similar to previous works, and thus we use ALE as a performance reference. We compare ALE, DARK-L, DARK-H, and DARK on four benchmark datasets. The ZSL ACC comparison is shown in Figure 2 and the GZSL H comparison is shown in Figure 3. From the results, we can observe the following interesting phenomenons.

Firstly, by comparing DARH-L and DARK-H to ALE, it can be noticed combining label-view ranking or hardness assessment individually with image-view ranking consistently leads to better performance. On the other hand, removing either one of them from DARK results in observable performance drop. Both observations clearly indicate that label-view ranking and hardness assessment are beneficial for learning good image-class similarity function for ZSL.

Secondly, we can observe that label-view ranking plays an important role in DARK. The main motivation of DARK is to consider the ranking information from both image view and label view. The results demonstrate that considering label view ranking contributes a lot to ZSL. In addition, if we only consider dual-view ranking and ignore hardness assessment, i.e., DARK-H, the average ACC and H are 60.73 and 32.85 respectively, which are better than the state-of-the-arts. The results validate that considering dual-view ranking is superior to only image-view ranking. In fact, ZSL is modeled as a cross-modality matching problem as in Eq. (1) where the knowledge from both modalities is useful (Karpathy, Joulin, and Li 2014). However, many previous ZSL approaches failed to take this important issue into account.

Thirdly, hardness assessment consistently improves the performance, which can be verified by comparing DARK-H to DARK or DARK-L to ALE. In fact, ZSL can be regarded as a special metric learning problem, where each class has only one label vector and many image vectors. Hard triplets contain important information but are more likely to be outliers of a class. If the attention is paid only to them, the major characteristics of a class may be ignored. In fact, as each class has only one label vector, it is more important to find the general information from images instead of some specific features from hard samples. On the other hand, if we treat them with the same weight the easy ones may dominate the loss function which is harmful for learning (Bishop and others 2006). We propose to assign weight to triplets based on Eq. (8) which is capable of taking both sides into account.

Conclusion

In this paper we consider ZSL problem. Many previous ZSL works learn the image-class similarity function by ranking based loss which aims to rank true label of an image before other labels. We argue that this image-view label ranking is not sufficient to construct an effective ZSL model and propose a novel dual-view ranking loss which further performs label-view image ranking by putting images belonging to a class before images from other classes using a point-to-set metric. In addition, previous works fail to well utilize the hardness of samples which either use only the hard ones or all samples with the same weight. We propose to perform a hardness assessment and then assign to triplets different

[1]DARK minus label-view ranking, analogous to DARK-H.

weights based on the current loss, resulting in a more accurate and robust model. Experiments on four benchmark datasets demonstrate that DARK significantly outperforms the state-for-the-art approaches for (generalized) ZSL.

Acknowledgments

This work was supported by the National Key R&D Program of China (No. 2018YFC0806900), the National Natural Science Foundation of China (No. 61571269, 61327902, 61525206, 61671196, 61701273), the National Postdoctoral Program for Innovative Talents (No. BX20180172), the Zhejiang Province Nature Science Foundation of China (No. LR17F030006), the China Postdoctoral Science Foundation (No. 2018T110100, 2017M610897), the Fundamental Research Funds for the Central Universities (Grant No. FRF-TP-18-016A1), and Berkeley DeepDrive.

References

Akata, Z.; Reed, S. E.; Walter, D.; Lee, H.; and Schiele, B. 2015. Evaluation of output embeddings for fine-grained image classification. In *CVPR*.

Akata, Z.; Perronnin, F.; Harchaoui, Z.; and Schmid, C. 2016. Label-embedding for image classification. *IEEE TPAMI*.

Annadani, Y., and Biswas, S. 2018. Preserving semantic relations for zero-shot learning. In *CVPR*.

Bishop, C. M., et al. 2006. *Pattern recognition and machine learning*, volume 1. Springer, New York.

Changpinyo, S.; Chao, W.; Gong, B.; and Sha, F. 2016. Synthesized classifiers for zero-shot learning. In *CVPR*.

Changpinyo, S.; Chao, W.; and Sha, F. 2017. Predicting visual exemplars of unseen classes for zero-shot learning. In *ICCV*.

Chao, W.; Changpinyo, S.; Gong, B.; and Sha, F. 2016. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *ECCV*, 52–68.

Farhadi, A.; Endres, I.; Hoiem, D.; and Forsyth, D. A. 2009. Describing objects by their attributes. In *CVPR*.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; and Mikolov, T. 2013. Devise: A deep visual-semantic embedding model. In *NIPS*.

Fu, Y.; Hospedales, T. M.; Xiang, T.; and Gong, S. 2015a. Transductive multi-view zero-shot learning. *IEEE TPAMI*.

Fu, Z.; Xiang, T.; Kodirov, E.; and Gong, S. 2015b. Zero-shot object recognition by semantic manifold distance. In *CVPR*.

Guo, Y.; Ding, G.; Jin, X.; and Wang, J. 2016. Transductive zero-shot recognition via shared model space learning. In *AAAI*.

Guo, Y.; Ding, G.; Han, J.; and Gao, Y. 2017a. Synthesizing samples for zero-shot learning. In *IJCAI*.

Guo, Y.; Ding, G.; Han, J.; and Gao, Y. 2017b. Zero-shot learning with transferred samples. *IEEE TIP* 26(7):3277–3290.

Guo, Y.; Ding, G.; Han, J.; Zhao, S.; and Wang, B. 2018. Implicit non-linear similarity scoring for recognizing unseen classes. In *IJCAI*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Jurie, F.; Bucher, M.; and Herbin, S. 2017. Generating visual representations for zero-shot classification. In *ICCV Workshops*.

Karpathy, A.; Joulin, A.; and Li, F. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*.

Kodirov, E.; Xiang, T.; Fu, Z.; and Gong, S. 2015. Unsupervised domain adaptation for zero-shot learning. In *ICCV*.

Kodirov, E.; Xiang, T.; and Gong, S. 2017. Semantic autoencoder for zero-shot learning. In *CVPR*.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2014. Attribute-based classification for zero-shot visual object categorization. *IEEE TPAMI*.

Li, Y.; Zhang, J.; Zhang, J.; and Huang, K. 2018. Discriminative learning of latent features for zero-shot recognition. In *CVPR*.

Liu, W.; Wen, Y.; Yu, Z.; and Yang, M. 2016. Large-margin softmax loss for convolutional neural networks. In *ICML*.

Morgado, P., and Vasconcelos, N. 2017. Semantically consistent regularization for zero-shot recognition. In *CVPR*.

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

Patterson, G., and Hays, J. 2012. SUN attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*.

Romera-Paredes, B., and Torr, P. H. S. 2015. An embarrassingly simple approach to zero-shot learning. In *ICML*.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *CVPR*.

Shrivastava, A.; Gupta, A.; and Girshick, R. B. 2016. Training region-based object detectors with online hard example mining. In *CVPR*.

Socher, R.; Ganjoo, M.; Manning, C. D.; and Ng, A. Y. 2013. Zero-shot learning through cross-modal transfer. In *NIPS*.

Socher, R.; Karpathy, A.; Le, Q. V.; Manning, C. D.; and Ng, A. Y. 2014. Grounded compositional semantics for finding and describing images with sentences. *TACL* 2:207–218.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Wu, F.; Tian, K.; Guan, J.; and Zhou, S. 2018. Global semantic consistency for zero-shot learning. In *ECCV*.

Xian, Y.; Akata, Z.; Sharma, G.; Nguyen, Q. N.; Hein, M.; and Schiele, B. 2016. Latent embeddings for zero-shot classification. In *CVPR*.

Xian, Y.; Lampert, C. H.; Schiele, B.; and Akata, Z. 2017. Zero-shot learning - A comprehensive evaluation of the good, the bad and the ugly. *CoRR* abs/1707.00600.

Xian, Y.; Schiele, B.; and Akata, Z. 2017. Zero-shot learning - the good, the bad and the ugly. In *CVPR*.

Zhang, H., and Koniusz, P. 2018. Zero-shot kernel learning. In *CVPR*.

Zhang, Z., and Saligrama, V. 2015. Zero-shot learning via semantic similarity embedding. In *ICCV*.

Zhou, S.; Wang, J.; Wang, J.; Gong, Y.; and Zheng, N. 2017. Point to set similarity based deep feature learning for person re-identification. In *CVPR*.

Zhu, Y.; Elhoseiny, M.; Liu, B.; and Elgammal, A. M. 2018. A generative adversarial approach for zero-shot learning from noisy texts. In *CVPR*.