Zero-Shot Learning - The Good, the Bad and the Ugly
Xian, Y.; Schiele, B.; Akata, Z.

Published in:
30th IEEE Conference on Computer Vision and Pattern Recognition

DOI:
10.1109/CVPR.2017.328

Citation for published version (APA):
Zero-Shot Learning - The Good, the Bad and the Ugly

Yongqin Xian\textsuperscript{1}  Bernt Schiele\textsuperscript{1}  Zeynep Akata\textsuperscript{1,2}

\textsuperscript{1}Max Planck Institute for Informatics  \textsuperscript{2}Amsterdam Machine Learning Lab
Saarland Informatics Campus  University of Amsterdam

Abstract

Due to the importance of zero-shot learning, the number of proposed approaches has increased steadily recently. We argue that it is time to take a step back and to analyze the status quo of the area. The purpose of this paper is threefold. First, given the fact that there is no agreed upon zero-shot learning benchmark, we first define a new benchmark by unifying both the evaluation protocols and data splits. This is an important contribution as published results are often not comparable and sometimes even flawed due to, e.g. pre-training on zero-shot test classes. Second, we compare and analyze a significant number of the state-of-the-art methods in depth, both in the classic zero-shot setting but also in the more realistic generalized zero-shot learning setting. Finally, we discuss limitations of the current status of the area which can be taken as a basis for advancing it.

1. Introduction

Zero-shot learning aims to recognize objects whose instances may not have been seen during training \cite{17,22,23,30,40}. The number of new zero-shot learning methods proposed every year has been increasing rapidly, i.e. the good aspects as our title suggests. Although each new method has been shown to make progress over the previous one, it is difficult to quantify this progress without an established evaluation protocol, i.e. the bad aspects. In fact, the quest for improving numbers has lead to even flawed evaluation protocols, i.e. the ugly aspects. Therefore, in this work, we propose to extensively evaluate a significant number of recent zero-shot learning methods in depth on several small to large-scale datasets using the same evaluation protocol both in zero-shot, i.e. training and test classes are disjoint, and the more realistic generalized zero-shot learning settings, i.e. training classes are present at test time.

We benchmark and systematically evaluate zero-shot learning w.r.t. three aspects, i.e. methods, datasets and evaluation protocol. The crux of the matter for all zero-shot learning methods is to associate observed and non-observed classes through some form of auxiliary information which encodes visually distinguishing properties of objects. Different flavors of zero-shot learning methods that we evaluate in this work are linear \cite{11,2,32} and nonlinear \cite{39,34} compatibility learning frameworks whereas an orthogonal direction is learning independent attribute \cite{22} classifiers and finally others \cite{42,7,26} propose a hybrid model between independent classifier learning and compatibility learning frameworks.

We thoroughly evaluate the second aspect of zero-shot learning, by using multiple splits of several small to large-scale datasets \cite{28,38,22,10,9}. We emphasize that it is hard to obtain labeled training data for fine-grained classes of rare objects recognizing which requires expert opinion. Therefore, we argue that zero-shot learning methods should be evaluated mainly on least populated or rare classes.

We propose a unified evaluation protocol to address the third aspect of zero-shot learning which is arguably the most important one. We emphasize the necessity of tuning hyperparameters of the methods on a validation class split that is disjoint from training classes as improving zero-shot learning performance via tuning parameters on test classes violates the zero-shot assumption. We argue that per-class averaged top-1 accuracy is an important evaluation metric when the dataset is not well balanced with respect to the number of images per class. We point out that extracting image features via a pre-trained deep neural network (DNN) on a large dataset that contains zero-shot test classes also violates the zero-shot learning idea as image feature extraction is a part of the training procedure. Moreover, we argue that demonstrating zero-shot performance on small-scale and coarse grained datasets, i.e. aPY \cite{10} is not conclusive. We recommend to abstract away from the restricted nature of zero-shot evaluation and make the task more practical by including training classes in the search space, i.e. generalized zero-shot learning setting. Therefore, we argue that our work plays an important role in advancing the zero-shot learning field by analyzing the good and bad aspects of the zero-shot learning task as well as proposing ways to eliminate the ugly ones.
2. Related Work

We review related work on zero-shot and generalized zero-shot learning, we present previous evaluations on the same task and describe the unique aspects of our work.

Zero-Shot Learning. In zero-shot learning setting test and training class sets are disjoint, which can be tackled by solving related sub-problems, e.g. learning intermediate attribute classifiers, and learning a mixture of seen class proportions, or by a direct approach, e.g. compatibility learning frameworks for being recent compatibility learning methods with similar loss functions; ESZSL for adding a regularization term to unregularized compatibility learning methods; and CMT proposing non-linear extensions to bilinear compatibility learning framework and finally SYNC for reporting the state-of-the-art on several benchmark datasets.

Generalized Zero-shot Learning. This setting generalizes the zero-shot learning task to the case with both seen and unseen classes at test time. Argues that although ImageNet classification challenge performance has reached beyond human performance, we do not observe similar behavior of the methods that compete at the detection challenge which involves rejecting unknown objects while detecting the position and label of a known object. Uses label embeddings to operate on the generalized zero-shot learning setting whereas proposes to learn latent representations for images and classes through coupled linear regression of factorized joint embeddings. On the other hand, introduces a new model layer to the deep net which estimates the probability of an input being from an unknown class and proposes a novelty detection mechanism. We evaluate and for being the most widely used.

Previous Evaluations of Zero-Shot Learning. In the literature some zero-shot vs generalized zero-shot learning evaluation works exist. Among these, proposes a model to learn the similarity between images and semantic embeddings on the ImageNet 1K by using 800 classes for training and 200 for test. Provides a comparison between five methods evaluated on three datasets including ImageNet with three standard splits and proposes a metric to evaluate generalized zero-shot learning performance.

Our work. We evaluate ten zero-shot learning methods on five datasets with several splits both for zero-shot and generalized zero-shot learning settings, provide statistical significance and robustness tests, and present other valuable insights that emerge from our benchmark. In this sense, ours is a more extensive evaluation compared to prior work.

3. Evaluated Methods

We start by formalizing the zero-shot learning task and then we describe the zero-shot learning methods that we evaluate in this work. Given a training set \( S = \{(x_n, y_n), n = 1...N\} \), with \( y_n \in \mathcal{Y}^{tr} \) belonging to training classes, the task is to learn \( f: \mathcal{X} \rightarrow \mathcal{Y} \) by minimizing the regularized empirical risk:

\[
\frac{1}{N} \sum_{n=1}^{N} L(y_n, f(x_n; W)) + \Omega(W)
\]

with \( L(.) \) being the loss function and \( \Omega(.) \) being the regularization term. Here, the mapping \( f: \mathcal{X} \rightarrow \mathcal{Y} \) from input to output embeddings is defined as:

\[
f(x; W) = \arg\max_{y \in \mathcal{Y}} F(x, y; W)
\]

At test time, in zero-shot learning setting, the aim is to assign a test image to an unseen class label, i.e. \( \mathcal{Y}^{ts} \subset \mathcal{Y} \) and in generalized zero-shot learning setting, the test image can be assigned either to seen or unseen classes, i.e. \( \mathcal{Y}^{tr+ts} \subset \mathcal{Y} \) with the highest compatibility score.

3.1. Learning Linear Compatibility

Attribute Label Embedding (ALE) use bi-linear compatibility function to associate visual and auxiliary information:

\[
F(x, y; W) = \theta(x)^TW\phi(y)
\]

where \( \theta(x) \) and \( \phi(y) \), i.e. image and class embeddings, both of which are given. \( F(.) \) is parameterized by the mapping \( W \), to be learned. Embarassingly Simple Zero Shot Learning (ESZSL) adds a regularization term to this objective. In the following, we provide a unified formulation of these four zero-shot learning methods.

DEVISE uses pairwise ranking objective that is inspired from unregularized ranking SVM:

\[
\sum_{y \in \mathcal{Y}^{tr}} [\Delta(y_n, y) + F(x_n, y; W) - F(x_n, y_n; W)]_+
\]

ALE uses weighted approximate ranking objective:

\[
\sum_{y \in \mathcal{Y}^{tr}} \frac{1}{\Delta(x_n, y_n)} [\Delta(y_n, y) + F(x_n, y; W) - F(x_n, y_n; W)]_+
\]
where \( l_k = \sum_{i=1}^{k} \alpha_i \) and \( r_{\Delta(x_n,y_n)} \) is defined as:

\[
\sum_{y \in Y^{tr}} 1(F(x_n,y;W) + \Delta(y_n,y) \geq F(x_n,y_n;W)) \tag{6}
\]

Following the heuristic in [13], [3] selects \( \alpha_i = 1/i \) which puts high emphasis on the top of the rank list.

SJE [4] gives full weight to the top of the ranked list and is inspired from the structured SVM [35]:

\[
[\max_{y \in Y^{tr}} (\Delta(y_n,y) + F(x_n,y;W)) - F(x_n,y_n;W)]_+ \tag{7}
\]

ESZSL [32] adds the following regularization term to the unregularized risk minimization formulation:

\[
\gamma \|W\phi(y)\|_F^2 + \lambda \|\theta(x)^TW\|_F^2 + \beta \|W\|_F^2 \tag{8}
\]

where \( \gamma, \lambda, \beta \) are parameters of this regularizer.

3.2. Learning Nonlinear Compatibility

Latent Embeddings (LATEM) [39] and Cross Modal Transfer (CMT) [34] encode an additional non-linearity in compatibility learning framework.

LATEM [39] constructs a piece-wise linear compatibility:

\[
F(x,y;W_i) = \max_{1 \leq i \leq K} \theta(x)^TW_i\phi(y) \tag{9}
\]

where every \( W_i \) models a different visual characteristic of the data and the selection of which matrix to use to do the mapping is a latent variable. LATEM uses the ranking loss formulated in \( \text{Equation 4} \)

CMT [34] first maps images into a semantic space of words, i.e. class names, where a neural network with \( \tanh \) non-linearity learns the mapping:

\[
\sum_{y \in Y^{tr}} \sum_{x \in X_y} \|\phi(y) - W_1 \tanh(W_2 \theta(x))\| \tag{10}
\]

where \((W_1, W_2)\) are weights of the two layer neural network. This is followed by a novelty detection mechanism that assigns images to unseen or seen classes. The novelty is detected either via thresholds learned using the embedded images of the seen classes or the outlier probabilities are obtained in an unsupervised way.

3.3. Learning Intermediate Attribute Classifiers

Although Direct Attribute Prediction (DAP) [22] has been shown to perform poorly compared to compatibility learning frameworks [3], we include it to our evaluation for being historically one of the most widely used methods in the literature.

DAP [22] learns probabilistic attribute classifiers and makes a class prediction by combining scores of the learned attribute classifiers. A novel image is assigned to one of the unknown classes using:

\[
f(x) = \arg\max_c \prod_{m=1}^{M} \frac{p(a_m^r|x)}{p(a_m^r)} \tag{11}
\]

with \( M \) being the total number of attributes. We train a one-vs-rest SVM with log loss that gives probability scores of attributes with respect to training classes.

3.4. Hybrid Models

Semantic Similarity Embedding (SSE) [42]. Convex Combination of Semantic Embeddings (CONSE) [26] and Synthesized Classifiers (SYNC) [7] express images and semantic class embeddings as a mixture of seen class proportions, hence we group them as hybrid models.

SSE [42] leverages similar class relationships both in image and semantic embedding space. An image is labeled with:

\[
\arg\max_{u \in U} \pi(\theta(x)^T\psi(y_u)) \tag{12}
\]

where \( \pi, \psi \) are mappings of class and image embeddings into a common space. Specifically, \( \psi \) is learned by sparse coding and \( \pi \) is by class dependent transformation.

CONSE [26] learns the probability of a training image belonging to a training class:

\[
f(x,t) = \arg\max_{y \in Y^{tr}} p_t(y|x) \tag{13}
\]

where \( y \) denotes the most likely training label \((t=1)\) for image \( x \). Combination of semantic embeddings \((s)\) is used to assign an unknown image to an unseen class:

\[
\frac{1}{Z} \sum_{i=1}^{T} p_t(f(x,t)|s) \tag{14}
\]

where \( Z = \sum_{i=1}^{T} p_t(f(x,t)|s) \), \( f(x,t) \) denotes the \( t \)-th most likely label for image \( x \) and \( T \) controls the maximum number of semantic embedding vectors.

SYNC [7] learns a mapping between the semantic class embedding space and a model space. In the model space, training classes and a set of phantom classes form a weighted bipartite graph. The objective is to minimize distortion error:

\[
\min_{w_c,v_r} \|w_c - \sum_{r=1}^{R} s_{cr}v_r\|_2^2 \tag{15}
\]

Semantic and model spaces are aligned by embedding real \((w_c)\) and phantom classes \((v_r)\) in the weighted graph \((s_{cr})\).
4. Datasets and Evaluation Protocol

In this section, we provide several components of previously used and our proposed zero-shot and generalized zero-shot learning evaluation protocols, e.g., datasets, image and class encodings and the evaluation protocol.

4.1. Dataset Statistics

Among the most widely used datasets for zero-shot learning, we select two coarse-grained, one small and one medium-scale, and two fine-grained, both medium-scale, datasets with attributes and one large-scale dataset without. Here, we consider between 10K and 1M images, and, between 100 and 1K classes as medium-scale.

Attribute Datasets. Statistics of the attribute datasets are presented in Table 1. Attribute Pascal and Yahoo (aPY) is a small-scale coarse-grained dataset with 64 attributes. Among the total number of 32 classes, 20 Pascal classes are used for training (we randomly select 5 for validation) and 12 Yahoo classes are used for testing. Animals with Attributes (AWA) is a coarse-grained dataset that is medium-scale in terms of the number of images, i.e., 30,475 and small-scale in terms of number of classes, i.e., 50. AWA introduces a standard zero-shot split with 40 classes for training (we randomly select 13 for validation) and 10 for testing. AWA has 85 attributes. Caltech-UCSD-Birds 200-2011 (CUB) is a fine-grained and medium-scale dataset with respect to both number of images and number of classes, i.e., 11,788 images from 200 different types of birds annotated with 312 attributes. SUN introduces the first zero-shot split of CUB with 150 training (50 validation classes) and 50 test classes. SUN is a fine-grained and medium-scale dataset with respect to both number of images and number of classes, i.e., SUN contains 14,340 images coming from 717 types of scenes annotated with 102 attributes. Following AWA we use 645 classes of SUN for training (we randomly select 65 for val) and 72 for testing.

Large-Scale ImageNet. We also evaluate the performance of methods on the large scale ImageNet. Among the total of 21K classes, 1K classes are used for training (we use 200 classes for validation) and the test split is either all the remaining 21K classes or a subset of it, e.g., we determine these subsets based on the hierarchical distance between classes and the population of classes.

4.2. Proposed Evaluation Protocol

We present our proposed unified protocol for image and class embeddings, dataset splits and evaluation criteria.

Image and Class Embedding. We extract image features from the entire image for SUN, CUB, AWA and ImageNet, with no image pre-processing. For aPY, as proposed in [10], we extract image features from bounding boxes. Our image embeddings are 2048-dim top-layer pooling units of the 101-layered ResNet as we found that it performs better than 1,024-dim top-layer pooling units of GoogleNet [35]. ResNet is pre-trained on ImageNet 1K and not fine-tuned. In addition to ResNet features, we evaluate methods with their published image features. As class embeddings, for aPY, AWA, CUB and SUN, we use per-class attributes. For ImageNet we use Word2Vec [25] provided by [7] as it does not contain attribute annotation for all the classes.

Dataset Splits. Zero-shot learning assumes disjoint training and test classes with the presence of all the images of training classes and the absence of any image from test classes during training. On the other hand, as deep neural network (DNN) training for image feature extraction is actually a part of model training, the dataset used to train DNNs, e.g. ImageNet, should not include any of the test classes. However, we notice from the standard splits (SS) of aPY and AWA datasets that 7 aPY test classes out of 12 (monkey, wolf, zebra, mug, building, bag, carriage), 6 AWA test classes out of 10 (chimpanzee, giant panda, leopard, persian cat, pig, hippopotamus), are among the 1K classes of ImageNet, i.e., are used to pre-train ResNet. On the other hand, the mostly widely used splits, i.e., we term them as standard splits (SS), for SUN from [22] and CUB from [2] shows us that 1 CUB test class out of 50 (Indigo Bunting),

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size</th>
<th>Detail</th>
<th>At</th>
<th>Y</th>
<th>(Y_{tr})</th>
<th>(Y_{ts})</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUN [28]</td>
<td>medium</td>
<td>fine</td>
<td>102</td>
<td>717</td>
<td>580 + 65</td>
<td>72</td>
</tr>
<tr>
<td>CUB [38]</td>
<td>medium</td>
<td>fine</td>
<td>312</td>
<td>200</td>
<td>100 + 50</td>
<td>50</td>
</tr>
<tr>
<td>AWA [22]</td>
<td>medium</td>
<td>coarse</td>
<td>85</td>
<td>50</td>
<td>27 + 13</td>
<td>10</td>
</tr>
<tr>
<td>aPY [10]</td>
<td>small</td>
<td>coarse</td>
<td>64</td>
<td>32</td>
<td>15 + 5</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 1: Statistics for attribute datasets: SUN [28], CUB [38], AWA [22], aPY [10] in terms of size of the datasets, fine-grained or coarse-grained, number of attributes, number of classes in training + validation \((Y_{tr})\) and test classes \((Y_{ts})\), number of images at training and test time for standard split (SS) and our proposed splits (PS).
and 6 SUN test classes out of 72 (restaurant, supermarket, planetarium, tent, market, bridge), are also among the 1K classes of ImageNet. We noticed that the accuracy for all methods on those overlapping test classes are higher than others. Therefore, we propose new dataset splits, i.e. proposed splits (PS), ensuring that none of the test classes appear in ImageNet 1K, i.e. used to train the ResNet model. We present the differences between the standard splits (SS) and the proposed splits (PS) in Table 1. While in SS and PS no image from test classes is present at training time, at test time SS does not include any images from training classes however our PS does. We designed the PS this way as evaluating accuracy on both training and test classes is crucial to show the generalization of methods.

ImageNet with thousands of classes provides possibilities of constructing several zero-shot evaluation splits. Following [7], our first two standard splits consider all the classes that are 2-hops and 3-hops away from the original 1K classes according to the ImageNet label hierarchy, corresponding to 1509 and 7678 classes. This split measures the generalization ability of the models with respect to the hierarchical and semantic similarity between classes. Our proposed split considers 500, 1K and 5K most populated classes among the remaining 21K classes of ImageNet with \( \approx 1756, \approx 1624 \) and \( \approx 1335 \) images per class on average. Similarly, we consider 500, 1K and 5K least populated classes in ImageNet which correspond to most fine-grained subsets of ImageNet with \( \approx 1, \approx 3 \) and \( \approx 51 \) images per class on average. Our final split considers all the remaining \( \approx 20K \) classes of ImageNet with at least 1 image per-class, \( \approx 631 \) images per class on average.

**Evaluation Criteria.** Single label image classification accuracy has been measured with Top-1 accuracy, i.e. the prediction is accurate when the predicted class is the correct one. If the accuracy is averaged for all images, high performance on densely populated classes is encouraged. However, we are interested in having high performance also on sparsely populated classes. Therefore, we average the correct predictions independently for each class before dividing their cumulative sum w.r.t the number of classes, i.e. we measure average per-class top-1 accuracy.

In generalized zero-shot learning setting, the search space at evaluation time is not restricted to only test classes, but includes also the training classes, hence this setting is more practical. As with our proposed split at test time we have access to some images from training classes, after having computed the average per-class top-1 accuracy on training and test classes, we compute the harmonic mean of training and test accuracies:

\[
H = 2 \times \frac{acc_{tr} \times acc_{ts}}{acc_{tr} + acc_{ts}} \quad (16)
\]

where \( acc_{tr} \) and \( acc_{ts} \) represent the accuracy of images from seen \( (Y^{tr}) \), and images from unseen \( (Y^{ts}) \) classes respectively. We choose harmonic mean as our evaluation criteria and not arithmetic mean because in arithmetic mean if the seen class accuracy is much higher, it effects the overall results significantly. Instead, our aim is high accuracy on both seen and unseen classes.

5. Experiments

We first provide zero-shot learning results on attribute datasets SUN, CUB, AWA and aPY and then on the large-scale ImageNet dataset. Finally, we present results for the generalized zero-shot learning setting.

5.1. Zero-Shot Learning Results

On attribute datasets, i.e. SUN, CUB, AWA and aPY, we first reproduce the results of each method using their evaluation protocol, then provide a unified evaluation protocol using the same train/val/test class splits, followed by our proposed train/val/test class splits. We also evaluate the robustness of the methods to parameter tuning and visualize the ranking of different methods. Finally, we evaluate the methods on the large-scale ImageNet dataset.

**Reproducing Results.** For sanity-check, we re-evaluate methods \([22, 42, 39, 4, 32, 7]\) using provided features and code. We chose SUN and AWA as two representative of fine-grained and non-fine-grained datasets having been widely used in the literature. We observe from the results in Table 2 that our reproduced results and the reported results of DAP and SYNC are identical to the reported numbers in their original publications. For LATEM, we obtain slightly different results which can be explained by the non-convexity and thus the sensitivity to initialization. Similarly for SJE random sampling in SGD might lead to slightly different results. ESZSL has some variance because its algorithm randomly picks a validation set during each run, which leads to different hyperparameters. Notable observations on SSE \([42]\) results are as follows. The published code has hard-coded hyperparameters operational on aPY, \footnote{\([32]\) has public code available, but is not evaluated on SUN or AWA.}

<table>
<thead>
<tr>
<th>Model</th>
<th>SUN</th>
<th>AWA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R  O</td>
<td>R  O</td>
</tr>
<tr>
<td>DAP</td>
<td>22.1</td>
<td>41.4</td>
</tr>
<tr>
<td>SSE</td>
<td>83.0</td>
<td>64.9</td>
</tr>
<tr>
<td>LATEM</td>
<td>–</td>
<td>71.2</td>
</tr>
<tr>
<td>SJE</td>
<td>–</td>
<td>67.2</td>
</tr>
<tr>
<td>ESZSL</td>
<td>64.3</td>
<td>49.3</td>
</tr>
<tr>
<td>SYNC</td>
<td>62.8</td>
<td>69.7</td>
</tr>
</tbody>
</table>

Table 2: Reproducing zero-shot results: O = Original results published in the paper, R = Reproduced using provided image features and code. We measure top-1 accuracy in %.
i.e. number of iterations, number of data points to train SVM, and one regularizer parameter $\gamma$ which lead to inferior results than the ones reported here, therefore we set these parameters on validation sets. On SUN, SSE uses 10 classes (instead of 72) and our results with validated parameters got an improvement of 0.5% that may be due to random sampling of training images. On AWA, our reproduced result being 64.9% is significantly lower than the reported result (76.3%). However, we could not reach the reported result even by tuning parameters on the test set, i.e. we obtain 73.8% in this case.

**Reproduced Results vs Standard Split (SS).** In addition to \cite{22,42,39,4,32,7,34}, we re-implement \cite{26,11,3} based on the original publications. We use train, validation, test splits as provided in Table 1 and report results on Table 3 with deep ResNet features. DAP \cite{22} uses hand-crafted image features and thus reproduced results with those features are significantly lower than the results with deep features (22.1% vs 38.9%). When we investigate the results in detail, we noticed two irregularities with reported results on SUN. First, SSE \cite{42} and ESZSL \cite{32} report results on a test split with 10 classes whereas the standard split of SUN contains 72 test classes (74.5% vs 54.5% with SSE \cite{42} and

<table>
<thead>
<tr>
<th>Method</th>
<th>SUN SS</th>
<th>CUB PS</th>
<th>AWA SS</th>
<th>aPY PS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAP \cite{22}</td>
<td>38.9</td>
<td>39.9</td>
<td>37.5</td>
<td>40.0</td>
</tr>
<tr>
<td>CONSE \cite{26}</td>
<td>44.2</td>
<td>38.8</td>
<td>36.7</td>
<td>34.3</td>
</tr>
<tr>
<td>CMT \cite{34}</td>
<td>41.9</td>
<td>39.9</td>
<td>37.3</td>
<td>34.6</td>
</tr>
<tr>
<td>SSE \cite{42}</td>
<td>54.5</td>
<td>51.5</td>
<td>43.7</td>
<td>43.9</td>
</tr>
<tr>
<td>LATEM \cite{39}</td>
<td>56.9</td>
<td>55.3</td>
<td>49.4</td>
<td>49.3</td>
</tr>
<tr>
<td>ALE \cite{3}</td>
<td>59.1</td>
<td>58.1</td>
<td>53.2</td>
<td>54.9</td>
</tr>
<tr>
<td>DEVISE \cite{11}</td>
<td>57.5</td>
<td>56.5</td>
<td>53.2</td>
<td>52.0</td>
</tr>
<tr>
<td>SJE \cite{4}</td>
<td>57.1</td>
<td>53.7</td>
<td>55.3</td>
<td>53.9</td>
</tr>
<tr>
<td>ESZSL \cite{32}</td>
<td>57.3</td>
<td>54.5</td>
<td>55.1</td>
<td>53.9</td>
</tr>
<tr>
<td>SYNC \cite{7}</td>
<td>59.1</td>
<td>56.3</td>
<td>54.1</td>
<td>55.6</td>
</tr>
</tbody>
</table>

Table 3: Zero-shot on SS = Standard Split, PS = Proposed Split using ResNet features (top-1 accuracy in %).

![Figure 1](#) Ranking 10 models by setting parameters on three validation splits on the standard (SS, left) and proposed (PS, right) setting. Element $(i, j)$ indicates number of times model $i$ ranks at $j$th over all $4 \times 3$ observations. Models are ordered by their mean rank (displayed in brackets).

Figure 1: Ranking 10 models by setting parameters on three validation splits on the standard (SS, left) and proposed (PS, right) setting. Element $(i, j)$ indicates number of times model $i$ ranks at $j$th over all $4 \times 3$ observations. Models are ordered by their mean rank (displayed in brackets).

**Robustness.** We evaluate robustness of 10 methods to parameters by setting them on 3 different validation splits while keeping the test split intact. We report results on SS (Figure 2, top) and PS (Figure 2, bottom). On SUN and CUB, the results are stable across methods and across splits. This is expected as these datasets have balanced number of images across classes and due to their fine-grained nature, the validation splits are similar. On the other hand, AWA and aPY being small and coarse-grained datasets have several issues. First, many of the test classes on AWA and aPY are included in ImageNet1K. Second, they are not well balanced, i.e. different validation class splits contain significantly different number of images. Third, the class embeddings are far from each other, i.e. objects are semantically different, therefore different validation splits learn a different mapping between images and classes.

**Visualizing Method Ranking.** We rank the 10 methods based on their per-class top-1 accuracy using the non-parametric Friedman test \cite{14}, which does not assume a distribution on performance but rather uses algorithm ranking. Each entry of the rank matrix in Figure 1 indicates the number of times the method is ranked at the first to tenth rank. We then compute the mean rank of each method and order them based on that. Our general observation is that the highest ranked method on the standard split (SS) is SYNC while on the proposed split (PS) it is ALE. These results...
5.2. Generalized Zero-Shot Learning Results

In real world applications, image classification systems do not have access to whether a novel image belongs to a seen or unseen class in advance. Hence, generalized zero-shot learning is interesting from a practical point of view. Here, we use same models trained on zero-shot learning setting on our proposed splits (PS). We evaluate performance on both $\mathcal{Y}^{tr}$ and $\mathcal{Y}^{ts}$, i.e. using held-out images from $\mathcal{Y}^{ts}$.

As shown in Table 3, generalized zero-shot results are significantly lower than zero-shot results as training classes are included in the search space. Another interesting observation is that compatibility learning frameworks, e.g. ALE, DEVISE, SJE, perform well on test classes. However, methods that learn independent attribute or object classifiers, e.g. DAP and CONSE, perform well on training classes. Due to this discrepancy, we evaluate the harmonic mean which takes a weighted average of training and test
class accuracy. H measure ranks ALE as the best performing method on SUN, CUB and AWA datasets whereas on aPY dataset CMT* performs the best. Note that CMT* has an integrated novelty detection phase for which the method receives another supervision signal determining if the image belongs to a train or a test class. As a summary, generalized zero-shot learning setting provides one more level of detail on the performance of zero-shot learning methods. Our take-home message is that the accuracy of training classes is as important as the accuracy of test classes in real world scenarios. Therefore, methods should be designed in a way that they are able to predict labels well in train and test classes.

Visualizing Method Ranking. Similar to the analysis in the previous section, we rank the 11 methods based on per-class top-1 accuracy on train classes, test classes and based on Harmonic mean of the two. Looking at the rank matrix obtained by evaluating on test classes, i.e. Figure 3 top left, highest ranked 5 methods are the same as in Figure 1 i.e. ALE, DEVISE, SJE, LATEM, ESZSL while overall the absolute numbers are lower. Looking at the rank matrix obtained by evaluating the harmonic mean, i.e. Figure 3 bottom, the highest ranked 3 methods are the same as in Figure 1 i.e. ALE, DEVISE, SJE. Looking at the rank matrix obtained by evaluating on train classes, i.e. Figure 3 top right, our observations are different from Figure 1. ALE is ranked the 3rd but other highest ranked methods are at the bottom of this rank list. These results clearly suggest that we should not only optimize for test class accuracy but also for train class accuracy when evaluating zero-shot learning. Our final observation from Figure 3 is that CMT* is better than CMT in all cases which supports the argument that a simple novelty detection scheme helps to improve results.

### 6. Conclusion

In this work, we evaluated a significant number of state-of-the-art zero-shot learning methods on several datasets within a unified evaluation protocol both in zero-shot and generalized zero-shot settings. Our evaluation showed that compatibility learning frameworks have an edge over learning independent object or attribute classifiers and also over hybrid models. We discovered that some standard zero-shot splits may treat feature learning disjoint from the training stage and accordingly proposed new dataset splits. Moreover, disjoint training and validation class split is a necessary component of parameter tuning in zero-shot learning setting. Including training classes in the search space while evaluating the methods, i.e. generalized zero-shot learning, provides an interesting playground for future research. In summary, our work extensively evaluated the good and bad aspects of zero-shot learning while sanitizing the ugly ones.

### Table 5: Generalized Zero-Shot Learning on Proposed Split (PS) measuring ts = Top-1 accuracy on $Y^{ts}$, tr=Top-1 accuracy on $Y^{tr+ts}$, H = harmonic mean (CMT*: CMT with novelty detection). We measure top-1 accuracy in %.

<table>
<thead>
<tr>
<th>Method</th>
<th>SUN</th>
<th>CUB</th>
<th>AWA</th>
<th>aPY</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ts</td>
<td>tr</td>
<td>H</td>
<td>ts</td>
</tr>
<tr>
<td>DAP</td>
<td>9.2</td>
<td>25.1</td>
<td>7.2</td>
<td>1.7</td>
</tr>
<tr>
<td>CONSE</td>
<td>6.8</td>
<td>39.9</td>
<td>11.6</td>
<td>1.6</td>
</tr>
<tr>
<td>CMT</td>
<td>8.1</td>
<td>21.8</td>
<td>11.8</td>
<td>7.2</td>
</tr>
<tr>
<td>CMT*</td>
<td>8.7</td>
<td>28.0</td>
<td>13.3</td>
<td>4.7</td>
</tr>
<tr>
<td>SSE</td>
<td>2.1</td>
<td>36.4</td>
<td>4.0</td>
<td>8.5</td>
</tr>
<tr>
<td>LATEM</td>
<td>14.7</td>
<td>28.8</td>
<td>19.5</td>
<td>15.2</td>
</tr>
<tr>
<td>ALE</td>
<td>21.8</td>
<td>33.1</td>
<td>26.3</td>
<td>23.7</td>
</tr>
<tr>
<td>DEVISE</td>
<td>16.9</td>
<td>27.4</td>
<td>20.9</td>
<td>23.8</td>
</tr>
<tr>
<td>SJE</td>
<td>14.7</td>
<td>30.5</td>
<td>19.8</td>
<td>23.5</td>
</tr>
<tr>
<td>ESZSL</td>
<td>11.0</td>
<td>27.9</td>
<td>15.8</td>
<td>12.6</td>
</tr>
<tr>
<td>SYNC</td>
<td>7.9</td>
<td>43.3</td>
<td>13.4</td>
<td>11.5</td>
</tr>
</tbody>
</table>

Figure 3: Ranking 11 models on the proposed split (PS) in generalized zero-shot learning setting. Top-Left: on unseen classes (ts) accuracy, Top-Right: on seen classes (tr) accuracy, Bottom: on harmonic mean (H).
References