Zero-Shot Object Detection by Hybrid Region Embedding

Berkan Demirel
berkan.demirel@metu.edu.tr

Ramazan Gokberk Cinbis
gcinbis@ceng.metu.edu.tr

Nazli Ikizler-Cinbis
nazli@cs.hacettepe.edu.tr

HAVELSAN Inc. & Middle East
Technical University
Ankara, Turkey

Middle East Technical University
Ankara, Turkey

Hacettepe University
Ankara, Turkey

Abstract

Object detection is considered as one of the most challenging problems in computer vision, since it requires correct prediction of both classes and locations of objects in images. In this study, we define a more difficult scenario, namely *zero-shot object detection (ZSD)* where no visual training data is available for some of the target object classes. We present a novel approach to tackle this ZSD problem, where a convex combination of embeddings are used in conjunction with a detection framework. For evaluation of ZSD methods, we propose a simple dataset constructed from Fashion-MNIST images and also a custom zero-shot split for the Pascal VOC detection challenge. The experimental results suggest that our method yields promising results for ZSD.

1   Introduction

Object detection is one of the most studied tasks in computer vision research. Previously, mainstream approaches provided only limited success despite the efforts in carefully crafting representations for object detection, *e.g.* [43]. More recently, however, convolutional neural network (ConvNet) based models have lead to great advances in detection speed and accuracy, *e.g.* [27, 33, 35].

While the state-of-the-art in object detection is undoubtedly impressive, object detectors still lack semantic scalability. As these approaches rely heavily on fully supervised training schemes, one needs to collect large amounts of images with bounding box annotations for each target class of interest. Due to its laborious nature, data annotation remains as a major bottleneck in semantically enriching and universalizing object detectors.

Zero-shot learning (ZSL) aims to minimize the annotation requirements by enabling recognition of unseen classes, *i.e.* those with no training examples. This is achieved by transferring knowledge from seen to unseen classes by means of auxiliary data, typically obtained easily from textual sources. Mainstream examples for such ZSL approaches include methods for mapping visual and textual information into a joint space [1, 2, 4], and, those that explicitly leverage text-driven similarities across classes [30].

The existing ZSL approaches, however, predominantly focus on classification problems. In this work, we extend this ZSL paradigm to object detection and focus on the *zero-shot detection* (ZSD) task. Here, the goal is to recognize and localize instances of object classes with no training examples, purely based on auxiliary information that describes the class characteristics. The main motivation for studying ZSD is the observation that in most applications of ZSL, such as robotics, accurate object localization is equally important as recognition.

Our ZSD approach builds on the adaptation and combination of two mainstream approaches in zero-shot image classification: (i) convex combination of class embeddings [30], and, label embedding based classification [40]. More specifically, we propose a hybrid model that consists of two components: the first component leverages the detection scores of a supervised object detector to embed image regions into a class embedding space. The second component, on the other hand, learns a direct mapping from region pixels to the space of class embeddings. Both of these region embeddings are then converted into region detection scores by comparing their similarities with true class embeddings. Finally, we construct our zero-shot detector by integrating these two components into the the fast object detection framework YOLO [33].

We note that both components of our approach essentially provide an embedding of a given test image. Our main motivation in using them together is to employ two complementary sources of information. In particular, while the former component can be interpreted as a semantic composition method guided by class detection scores, the latter one focuses on transformation of image content into the class embedding space. Therefore, these two components are expected to better utilize semantic relations and visual cues, respectively.

In order to evaluate the effectiveness of the proposed ZSD approach, we define new benchmarks based on existing datasets. First, we create a simple ZSD dataset by composing images with multiple Fashion-MNIST [41] objects. Moreover, the Pascal VOC [12] dataset is similarly adapted to the ZSD task by defining new splits and settings. The experimental results show that our hybrid embedding approach yields promising results in both datasets.

To sum up, our main contributions in this work are as follows: (i) we define a novel zero-shot setting for detecting objects of unseen classes, (ii) we propose a novel hybrid method to handle newly defined ZSD task, (iii) we introduce two new benchmarks for evaluating ZSD approaches based on Fashion-MNIST and VOC datasets.

2   Related Work

In this section, we review related work on object detection and zero-shot learning.

Object Detection. State-of-the-art techniques in object detection are now all based on ConvNets. Among them, a well-known approach is the OverFeat framework [38]. OverFeat is essentially a fully-convolutional network that jointly makes predictions for object classification, localization and detection. Girshick *et al.* [16] propose R-CNN that extracts ConvNet features of pre-extracted object proposals. He *et al.* [18] and Girshick *et al.* [15] propose similar techniques that allow efficient feature extraction for object proposals, which greatly increases computational efficiency of the R-CNN-like approaches. Ren *et al.* [35] propose a further improved architecture that jointly learns to generate and score object proposals.

Redmon *et al.* [33, 34] propose YOLO (You Only Look Once) architecture that also simultenously generate candidate detection windows and per-window class probabilities. However, unlike R-CNN-like approaches, YOLO aims to produce all detections via a single
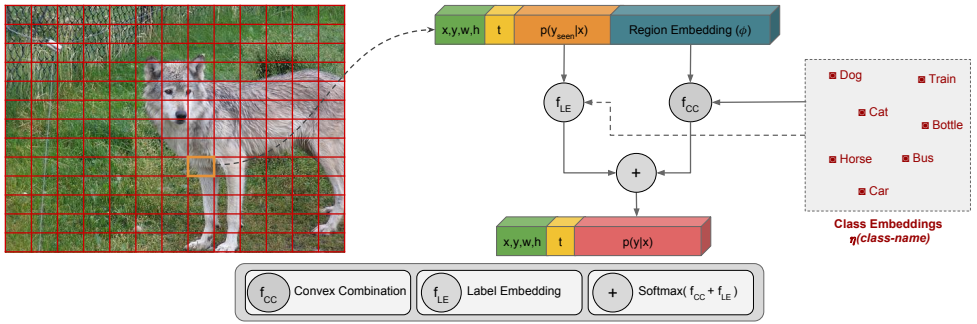
Figure 1: The framework of our ZSD model. In this model, (x, y, h, w) represents bounding box regression coordinates, t represents bounding box confidence score, $p(y_{seen}|c)$ represents initial class scores, ϕ represents embedding vector of related region, and $p(y|x)$ represents the final zero-shot detection class probabilities.

feed-forward pass of a ConvNet, avoiding need to process candidate windows in a sequential manner. *et al.* [27] use similar idea via multi-scale feature maps, and yield detection results using convolutional filters. Additional approaches in this context include using contextual and multi-scale representations [6], iterative grid without object proposals [29], using feature pyramids [26], or, landmark based detection [21].

Zero-Shot Learning. Zero-shot learning is the process of transferring recognition knowledge from seen classes to unseen classes. Most ZSL approaches use prior information from some sources to encode and transfer necessary knowledge. Attributes are one of the most important source of prior information [7, 17, 24, 25, 28]. Using attributes at different abstraction levels and propagating on different levels provide more distinctive information for unseen classes [3]. Other zero-shot learning methods are proposed, but in essence they commonly handle attributes, classes, features and their relations. Example studies use semantic class taxonomies [36], hierarchy and exclusion (HEX) graphs [9], random forests [22], linear layered networks [37], semantic autoencoders [23], visually meaningful word vectors [7], semantic dictionary [10] or manifold regularizations [42].

In recent years, label embedding methods have attracted attention in the zero-shot learning field. In this context, Akata *et al.* [1, 2] modify WSABIE formula [40] and use distributed word representation as side information. Frome *et al.* [14] use convolutional neural network architectures for mapping visual features into a rich semantic embedding space. Song *et al.* [39] propose a transductive learning (QFSL) method to learn unbiased embedding space since embedding spaces often have strong bias problem. Besides, visual-semantic discrepancy problem is observed when using textual side information. In this context, Demirel *et al.* [7] use attribute information as an intermediate layer to learn more generalizable distributed word representations.

Norouzi *et al.* [30] use convex combination of the semantic embedding vectors directly without learning any semantic space. Elhoseiny *et al.* [11] handle ZSL problem with purely textual descriptions. They define a constrained optimization formula that combine regression and knowledge transfer functions with additional constraints. Ba *et al.* [4] use MLP in their text pipeline to learn classifier weights of CNN in the image pipeline to handle zero-shot fine-

grained object classification. The defined MLP network generates a list of pseudo-attributes for each visual category by utilizing raw texts acquired from Wikipedia articles.

Finally, we note that while there are methods [19, 20, 53] for constructing object detectors through classifier-to-detector transfer, they still require carefully selected and annotated training images to train source classification models. This is fundamentally different from the ZSD task, which requires zero examples for unseen classes. By definition, therefore, ZSD is much more scalable towards large-scale recognition, compared to classifier-to-detector transfer approaches.

Recent works on ZSD. We note that unpublished manuscripts of three independent works on zero-shot detection have very recently appeared on arXiv [5, 52, 44]. Despite the common theme of zero-shot detection, these works significantly differ from our hybrid model in the following ways: [5] focuses on developing background-aware models, [52] proposes a semantic clustering loss, and [44] uses attribute based models. In addition, the experimental setup and datasets differ in all these works. We plan to investigate similarities and differences across these recent works and ours in future work.

3   Method

Our method consists of two components that (i) utilize a convex combination of class embeddings, an adaptation of the ideas from [30], and, (ii) directly learn to map regions to the space of class embeddings, by extending the label embedding approaches from zero-shot image classification [2].

The rest of this section explains the model details: in the first two sub-sections, we describe the convex combination and label embedding components. Then, we describe how we construct our zero-shot object detector within the YOLO detection framework.

3.1   Region Scoring by Convex Combination of Class Embeddings

The first component of our ZSD approach aims to semantically represent an image in the space of word vectors. More specifically, we represent a given image region (*i.e.* a bounding box) as the convex combination of training class embeddings, weighted by the class scores given by a supervised object detector of seen classes. The resulting semantic representation of the region is then utilized to estimate confidence scores for unseen classes.

This approach can be specified as follows: let $\mathcal{Y}_s$ be the set of seen classes, for which we have training images with bounding box annotations, and and let $\mathcal{Y}_u$ be the set of unseen classes, for which we have no visual training examples. Our goal is to learn a scoring function $f_{CC}(x,b,y) : \mathcal{X} \times \mathcal{B} \times \mathcal{Y} \to \mathcal{R}$ that measures the relevance of label $y \in \mathcal{Y}_s$, which can be a seen or unseen class, for a given candidate bounding box $b \in \mathcal{B}$ and the image $x \in \mathcal{X}$.

We assume that a d_e dimensional embedding vector $\eta(y)$, such as word embeddings of class names or class-wise attribute indicator vectors, is available for each class. The scoring function $f_{CC}(x,b,y)$ is then defined as the cosine similarity between the class embedding $\eta(y)$ and the image region embedding $\phi_{CC}(x,b)$:

$$f_{CC}(x,b,y) = \frac{\phi_{CC}(x,b)^{\mathrm{T}} \eta(y)}{\|\phi_{CC}(x,b)\| \|\eta(y)\|} \tag{1}$$

where $\phi_{CC}(x,b)$ is defined as follows:

$$\phi_{CC}(x,b) = \frac{1}{\sum_{y \in \mathcal{Y}_s} p(y|x,b)} \sum_{y \in \mathcal{Y}_s} p(y|x,b)\eta(y) \tag{2}$$

Here, $p(y|x,b)$ is the class posterior probability given by the supervised object detection model. Therefore, ϕ_{CC} can simply be interpreted as a weighted sum of class embeddings, over the seen classes.

3.2 Region Scoring by Label Embedding

The convex combination driven scoring function fcc utilizes detection scores and embeddings of the training classes to estimate scores of zero-shot classes. In the label embedding approach, however, our goal is to directly model the compatibility between the visual features of image regions and class embeddings. For this purpose, we define the label embedding driven scoring function $f_{LE}(x,b,y) : \mathcal{X} \times \mathcal{B} \times \mathcal{Y} \to \mathcal{R}$ that measures the relevance of label $y \in \mathcal{Y}$ for a given candidate bounding box $b \in \mathcal{B}$ in an image $x \in \mathcal{X}$ as follows:

$$f_{LE}(x,b,y) = \frac{\phi_{LE}(x,b)^{\mathrm{T}}\eta(y)}{\|\phi_{LE}(x,b)\|\|\eta(y)\|} \tag{3}$$

where $\phi_{LE}(x,b)$ is basically a deep convolutional neural network that maps the image region b of image x to the space of class embeddings.

We note that $f_{LE}(x,b,y)$ can equivalently be interpreted as a dot product between ℓ_2-normalized image region descriptors and class embeddings. While it is common to ℓ_2-normalize class embeddings in zero-shot image classification studies [2], we also ℓ_2-normalize the image embedding vectors. In our preliminary experiments, we have observed that this additional normalization step is beneficial for the zero-shot detection task.

We learn the $\phi_{LE}(x,b)$ network in an end-to-end fashion within our YOLO-based zero-shot detection framework, which we explain in the next section.

3.3 Zero-Shot Object Detection

We use the YOLO-v2 [33] architecture to construct our zero-shot object detector. The original YOLO architecture that we utilize contains a convolutional network that reduces the spatial dimensions of the input by a factor of 32 and results in a tensor of depth $k(5+|\mathcal{Y}_s|)$, *e.g.* an input image of size $416 \times 416 \times 3$ results in a tensor of size $13 \times 13 \times k(5+|\mathcal{Y}_s|)$. Each cell within this output tensor encodes the k detections per cell ($k = 5$ by default), and, each block of size $5+|\mathcal{Y}_s|$ encodes one such detection. Here, for a single detection, the first 4 dimensions encode the relative bounding box coordinates, the following dimension encodes the estimated window objectness score, and the final $|\mathcal{Y}_s|$ dimensions encode class confidence scores.

To adapt YOLO architecture for the zero-shot detection task, we modify it in the following manner: we increase the final output depth from $k(5+|\mathcal{Y}_s|)$ to $k(5+|\mathcal{Y}_s|+d_e)$, where the newly added d_e dimensions per detection correspond to the $\phi_{LE}(x,b)$ output of the label embedding component of the model. In this way, the same convolutional network is shared for candidate box prediction, class prediction and class-embedding prediction purposes.

During training, the original YOLO formulation uses three separate mean-squared error based loss functions, defined over the differences between predictions and ground truth values for (i) bounding boxes, (ii) intersection-over-union values, and, (iii) classes. For training f_{CC} defined in Eq. (1), the original YOLO loss function over class predictions is used as is. For training f_{LE} defined in Eq. (3), however, we extend the loss function by incorporating an additional loss function L_{LE}. L_{LE} basically measures correctness of the label embedding driven class predictions in a max-margin sense:

$$L_{LE}(x,b,y) = \frac{1}{|\mathcal{Y}_s|-1} \sum_{y' \in \mathcal{Y}_s \setminus \{y\}} \max\left(0, 1 - f_{LE}(x,b,y) + f_{LE}(x,b,y')\right) \tag{4}$$

where y is the ground-truth class corresponding to the bounding box b in input image x. Here, the goal is to ensure that at each window prediction, the label embedding based confidence score f_{LE} for the target class is larger than that of each other class. Other than this extension, we use the original YOLO training procedure, over the seen classes.

Once the network is trained, we jointly utilize the scoring functions f_{CC} and f_{LE} by computing soft-max of their summations, over the classes of interest:

$$p(y|x,b) = \frac{\exp\left(f_{CC}(x,b,y) + f_{LE}(x,b,y)\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(f_{CC}(x,b,y') + f_{LE}(x,b,y')\right)} \tag{5}$$

where $p(y|x,b)$ is the predicted posterior probability of (seen or unseen) class y given region b of image x. The final set of detections are obtained by using the non-maxima suppression procedure of YOLO over all candidate detection windows, objectness scores, and the final probabilities $p(y|x,b)$.

4   Experiments

In this section, we present our experimental evaluation of the proposed approach. In Section 4.1, we describe the ZSD datasets that we prepare and utilize. In Section 4.2, we explain class embeddings used in our experiments. Finally, in Section 4.3 and Section 4.4, we give the implementation details and our experimental results.

4.1   Datasets

We use two different datasets: Fashion-ZSD and Pascal-ZSD. We propose two new testbeds for evaluation of ZSD approaches. First, we create a synthetic dataset based on combinations of objects from the Fashion-MNIST [41] dataset. Second, we compose a new split based on existing Pascal VOC [12] benchmarks. The details of these testbeds are described below.

Fashion-ZSD: This is a toy dataset that we generate for evaluation of ZSD methods, based on the Fashion-MNIST [41] dataset. Fashion-MNIST originally consists of Zalando's article images with associated labels. This dataset contains 70,000 grayscale images of size 28x28, and 10 classes. For ZSD task, we split the dataset into two disjoint sets; seven classes are used in training and three classes are used as the unseen test classes (Table 1). We generate multi-object images such that there are three different objects in each image. Randomly cropped objects are utilized to create clutter regions. As shown in Figure 2, we consider four scenarios: no noise or occlusion, scenes with partial occlusions, those with clutter,

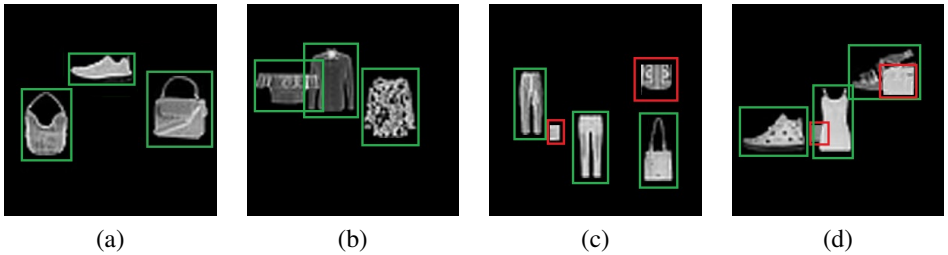| (a) | (b) | (c) | (d) |

Figure 2: Ground truth object regions are shown with green and noise regions are shown in red boxes. The dataset consists of images from four different scenarios. From left-to-right, (a) full objects only, (b) partial occlusions, (c) clutter regions included, and (d) a scene with both partial occlusions and clutter regions.

and, finally scenes with both partial occlusions and clutter regions. 8000 images of the resulting 16333 training images are held out for validation purposes. As a result, we obtain the Fashion-ZSD dataset with 8333 training, 8000 validation and 6999 test images.

Pascal-ZSD. This is an adapted version of the Pascal VOC datasets [12]. We select 16 of the 20 classes for training and the remaining 4 classes (*i.e.* car, dog, sofa and train) for test. The *train+val* subsets of Pascal VOC 2007 and 2012 datasets are used for training classes, and the *test* subset of Pascal 2007 is used for evaluation on the unseen classes. Images containing a mixture of train and test classes are ignored.

4.2 Class Embeddings

For the Fashion-ZSD dataset, we generate 300-dimensional GloVe word embedding vectors [31] for each class name, using Common Crawl Data. For the class names that contain multiple words, we take the average of the word vectors. For Pascal-ZSD, we use attribute annotations of aPaY dataset [13], since aPascal(aP) part of this dataset is obtained from Pascal VOC images. We average 64-dimensional indicator vectors of per-object attributes over the dataset to obtain class embeddings.

4.3 Zero-Shot Detection on Fashion-ZSD Dataset

In this part, we explain our ZSD experiments on Fashion-ZSD dataset. We initialize the convolutional layers of our model using the weights pre-trained on the ILSRVC12 [8] classification images. Training of our approach is completed in 10 epochs, where batch size is 32 and learning rate is 0.001. In our experiments, we first evaluate the performance of the trained network on seen training classes. According to the results presented in Table 1, the proposed approach obtains 91.9% mAP on the validation images with seen classes, which shows the proper training of the detection model. On the test set with unseen classes only, our proposed approach yields an mAP of 64.9%, highlighting the difficulty of zero-shot detection task even in simple settings.

On the combinated validation and test evaluation, our method achieves 81.7% mAP. This setting is particularly interesting, as it requires recognition over both seen and unseen objects at detection time. Our result suggests that the model is able to detect objects of unseen test classes even in the presence of seen classes, without being dominated by them.

Test split	Training Classes							Test Classes			mAP (%)
	t-shirt	trouser	coat	sandal	shirt	sneaker	bag	pullover	dress	ankle-boot	
val	.89	.91	.90	.97	.86	.99	.90	-	-	-	91.9
test	-	-	-	-	-	-	-	.49	.49	.95	64.9
val+test	.89	.90	.90	.97	.86	.99	.91	.45	.40	.90	81.7

Table 1: ZSD performances of proposed hybrid method on Fashion-ZSD dataset. We report class based average precision and mean average precision (mAP) scores.
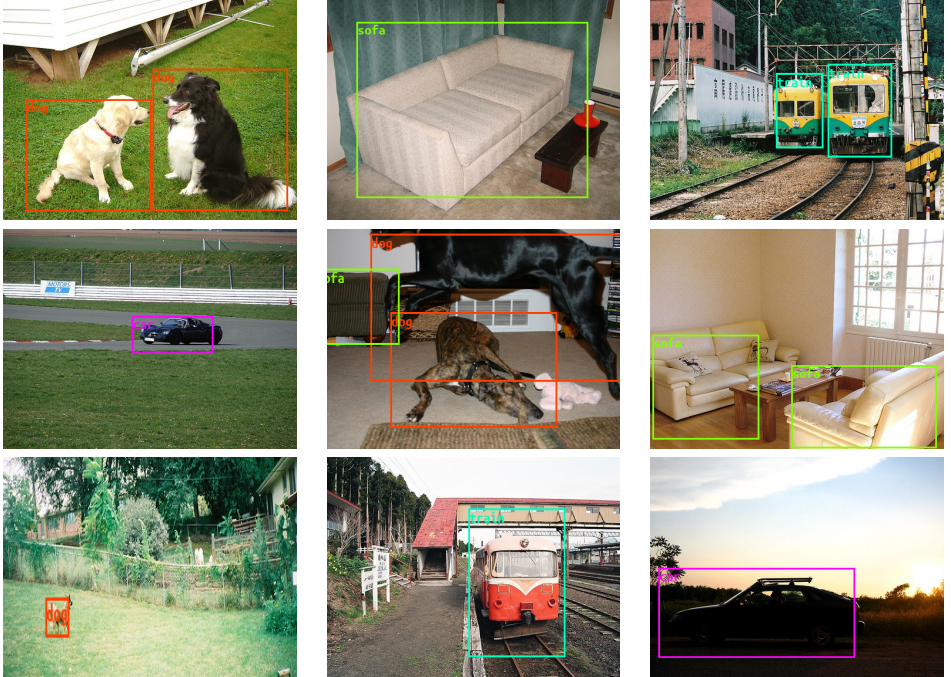

Figure 3: Successful detection results of unseen objects on Pascal-ZSD dataset using proposed hybrid region embedding.

4.4   Zero-Shot Detection on Pascal-ZSD Dataset

In this part, we explain our ZSD experiments on Pascal-ZSD dataset. Training settings of the proposed method on Pascal-ZSD dataset are same with the previous experiment, except that the number of epochs is set to 30. We present the results our approach, as well as individual performances of convex combination and label embedding components, in Table 2. The proposed hybrid approach yields 65.6% mAP on seen classes, 54.6% mAP on unseen classes and 52.3% mAP on the combination of seen and unseen classes. By comparing individual components of the model, we observe that convex combination (CC) outperforms label embedding (LE), and the hybrid scheme further improves the results.

The reason why the performance of the individual label embedding component is much lower can potentially be explained by the fact that the ZSD-Pascal dataset is relatively small: there are 16 classes in the training set, and this number is most probably insufficient to learn a direct mapping from visual features to class embeddings.

Qualitative results for our approach are provided in Figure 3. In this figure, example

Method	Test split	aeroplane	bicycle	bird	boat	bottle	bus	cat	chair	cow	dining table	horse	motorbike	person	potted plant	sheep	tvmonitor	car	dog	sofa	train	mAP (%)
LE	v	.46	.50	.44	.28	.12	.59	.44	.20	.11	.38	.35	.47	.65	.16	.18	.53	-	-	-	-	36.8
	t	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	.54	.79	.45	.12	47.9
	v+t	.34	.48	.40	.23	.12	.34	.28	.12	.09	.32	.28	.36	.60	.15	.13	.50	.27	.26	.20	.05	27.4
CC	v	.69	.74	.72	.63	.43	.83	.73	.43	.43	.66	.78	.80	.75	.41	.62	.75	-	-	-	-	65.0
	t	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	.60	.85	.44	.27	53.8
	v+t	.67	.73	.70	.59	.41	.61	.58	.32	.32	.65	.74	.68	.72	.39	.57	.72	.49	.24	.10	.15	52.0
H	v	.70	.73	.76	.54	.42	.86	.64	.40	.54	.75	.80	.80	.75	.34	.69	.79	-	-	-	-	**65.6**
	t	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	.55	.82	.55	.26	**54.2**
	v+t	.68	.72	.74	.48	.41	.61	.48	.25	.48	.73	.75	.71	.73	.33	.59	.57	.44	.25	.18	.15	**52.3**

Table 2: ZSD performances of proposed label embedding (LE), convex combination (CC) and hybrid (H) methods on Pascal-ZSD dataset. We report per-class average precision (AP) and mean-AP (mAP) scores of validation (v), test (t) and validation+test (v+t) images.

Figure 4: Unsuccessful detection results of unseen objects on Pascal-ZSD dataset using hybrid region embedding.

results of succesful detections of objects of unseen classes with various poses and sizes are shown. Additionally, example failure cases are shown on Figure 4. Problems in detection include missed detections, false positives, as well as misclassification of objects despite correct localization. For instance, in the second image within Figure 4, we see that *"picnic bench"* object is misrecognized as *"sofa"*, most probably due to relative similarity of the *'"chair"* and *"dining table"* seen classes in the embedding space.

5 Conclusion

Localization of instances of unseen classes is as important as their recognition in various applications, such as robotics. Moreover, to overcome the annotation bottleneck, alternative methods for training object detectors are needed. To this end, in this work, we handle the problem of zero-shot detection and propose a novel hybrid method that aggregates both label embeddings and convex combinations of semantic embeddings together in a region embedding framework. By integrating these two components within an object detector backbone, detection of classes with no visual examples becomes possible. We introduce two new testbeds for evaluating ZSD approaches, and our experimental results indicate that the proposed hybrid framework is a promising step towards achieving ZSD goals.

6 Acknowledgements

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 819–826. IEEE, 2013.

[2] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.* IEEE Computer Society, 2015.

[3] Ziad Al-Halah and Rainer Stiefelhagen. How to transfer? zero-shot object recognition via hierarchical transfer of semantic attributes. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pages 837–843. IEEE, 2015.

[4] Jimmy Ba, Kevin Swersky, Sanja Fidler, and Ruslan Salakhutdinov. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *Proc. IEEE Int. Conf. on Computer Vision*, 2015.

[5] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. *arXiv preprint arXiv:1804.04340*, 2018.

[6] Sean Bell, C Lawrence Zitnick, Kavita Bala, and Ross Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2874–2883, 2016.

[7] Berkan Demirel, Ramazan Gokberk Cinbis, and Nazli Ikizler-Cinbis. Attributes2classname: A discriminative model for attribute-based unsupervised zero-shot learning. In *Proc. IEEE Int. Conf. on Computer Vision*, Oct 2017.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 248–255. IEEE, 2009.

[9] Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In *Proc. European Conf. on Computer Vision*, pages 48–64. Springer, 2014.

[10] Zhengming Ding, Ming Shao, and Yun Fu. Low-rank embedded ensemble semantic dictionary for zero-shot learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2050–2058, 2017.

[11] Mohamed Elhoseiny, Babak Saleh, and Ahmed Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *Proc. IEEE Int. Conf. on Computer Vision*, pages 2584–2591, 2013.

[12] Mark Everingham, Luc Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *Int. J. on Computer Vision*, 88(2):303–338, 2010.

[13] Alireza Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1778–1785. IEEE, 2009.

[14] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 2121–2129, 2013.

[15] Ross Girshick. Fast r-cnn. In *Proc. IEEE Int. Conf. on Computer Vision*, pages 1440–1448, 2015.

[16] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 580–587, 2014.

[17] Yuchen Guo, Guiguang Ding, Jungong Han, and Sheng Tang. Zero-shot learning with attribute selection. *Entropy*, 20(40):60, 2018.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(9):1904–1916, 2015.

[19] Judy Hoffman, Sergio Guadarrama, Eric S Tzeng, Ronghang Hu, Jeff Donahue, Ross Girshick, Trevor Darrell, and Kate Saenko. Lsda: Large scale detection through adaptation. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 3536–3544, 2014.

[20] Judy Hoffman, Deepak Pathak, Trevor Darrell, and Kate Saenko. Detector discovery in the wild: Joint multiple instance and representation learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2883–2891, 2015.

[21] Lichao Huang, Yi Yang, Yafeng Deng, and Yinan Yu. Densebox: Unifying landmark localization with end to end object detection. *arXiv preprint arXiv:1509.04874*, 2015.

[22] Dinesh Jayaraman and Kristen Grauman. Zero-shot recognition with unreliable attributes. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 3464–3472, 2014.

[23] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3174–3183, 2017.

[24] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 951–958, 2009.

[25] C.H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(3):453–465, March 2014.

[26] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[27] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Proc. European Conf. on Computer Vision*, pages 21–37. Springer, 2016.

[28] Changzhi Luo, Zhetao Li, Kaizhu Huang, Jiashi Feng, and Meng Wang. Zero-shot learning via attribute regression and class prototype rectification. *IEEE Trans. on Image Processing*, 27(2):637–648, 2018.

[29] Mahyar Najibi, Mohammad Rastegari, and Larry S Davis. G-cnn: an iterative grid based object detector. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2369–2377, 2016.

[30] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013.

[31] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[32] Shafin Rahman, Salman Khan, and Fatih Porikli. Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. *arXiv preprint arXiv:1803.06049*, 2018.

[33] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7263–7271, 2017.

[34] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 779–788, 2016.

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 91–99, 2015.

[36] Marcus Rohrbach, Michael Stark, and Bernt Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1641–1648. IEEE, 2011.

[37] Bernardino Romera-Paredes and PHS Torr. An embarrassingly simple approach to zero-shot learning. In *Proc. Int. Conf. Mach. Learn.*, pages 2152–2161, 2015.

[38] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Robert Fergus, and Yann Lecun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *Proc. Int. Conf. Learn. Represent.*, 2014.

[39] Jie Song, Chengchao Shen, Yezhou Yang, Yang Liu, and Mingli Song. Transductive unbiased embedding for zero-shot learning. *arXiv preprint arXiv:1803.11320*, 2018.

[40] Jason Weston, Samy Bengio, and Nicolas Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine learning*, 81(1):21–35, 2010.

[41] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[42] Xing Xu, Fumin Shen, Yang Yang, Dongxiang Zhang, Heng Tao Shen, and Jingkuan Song. Matrix tri-factorization with manifold regularizations for zero-shot learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3798–3807, 2017.

[43] Junjie Yan, Zhen Lei, Longyin Wen, and Stan Li. The fastest deformable part model for object detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2497–2504, 2014.

[44] Pengkai Zhu, Hanxiao Wang, Tolga Bolukbasi, and Venkatesh Saligrama. Zero-shot detection. *arXiv preprint arXiv:1803.07113*, 2018.