Semi-supervised Vocabulary-informed Learning

Yanwei Fu and Leonid Sigal
Disney Research
y.fu@qmul.ac.uk, lsigal@disneyresearch.com

Abstract

Despite significant progress in object categorization, in recent years, a number of important challenges remain; mainly, ability to learn from limited labeled data and ability to recognize object classes within large, potentially open, set of labels. Zero-shot learning is one way of addressing these challenges, but it has only been shown to work with limited sized class vocabularies and typically requires separation between supervised and unsupervised classes, allowing former to inform the latter but not vice versa. We propose the notion of semi-supervised vocabulary-informed learning to alleviate the above mentioned challenges and address problems of supervised, zero-shot and open set recognition using a unified framework. Specifically, we propose a maximum margin framework for semantic manifold-based recognition that incorporates distance constraints from (both supervised and unsupervised) vocabulary atoms, ensuring that labeled samples are projected closest to their correct prototypes, in the embedding space, than to others. We show that resulting model shows improvements in supervised, zero-shot, and large open set recognition, with up to 310K class vocabulary on AwA and ImageNet datasets.

1. Introduction

Object recognition, and more specifically object categorization, has seen unprecedented advances in recent years with development of convolutional neural networks (CNNs) [23]. However, most successful recognition models, to date, are formulated as supervised learning problems, in many cases requiring hundreds, if not thousands, labeled instances to learn a given concept class [10]. This exuberant need for large labeled datasets has limited recognition models to domains with 100's to few 1000's of classes. Humans, on the other hand, are able to distinguish beyond $30,000$ basic level categories [5]. What is more impressive, is the fact that humans can learn from few examples, by effectively leveraging information from other object category classes, and even recognize objects without ever seeing them (*e.g.*, by reading about them on the Internet). This ability has spawned research in few-shot and zero-shot learning.

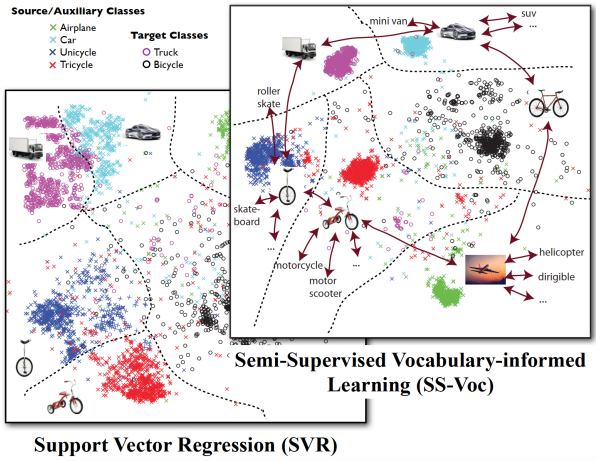

Figure 1. **Illustration of the semantic embeddings** learned (left) using support vector regression (**SVR**) and (right) using the proposed semi-supervised vocabulary-informed (**SS-Voc**) approach. In both cases, t-SNE visualization is used to illustrate samples from 4 source/auxiliary classes (denoted by $\times$) and 2 target/zero-shot classed (denoted by $\circ$) from the ImageNet dataset. Decision boundaries, illustrated by dashed lines, are drawn by hand for visualization. Note, that (i) large margin constraints in SS-Voc, both among the source/target classes and the external vocabulary atoms (denoted by arrows and words), and (ii) fine-tuning of the semantic word space, lead to a better embedding with more compact and separated classes (*e.g.*, see *truck* and *car* or *unicycle* and *tricycle*).

Zero-shot learning (ZSL) has now been widely studied in a variety of research areas including neural decoding by fMRI images [31], character recognition [26], face verification [24], object recognition [25], and video understanding [17, 45]. Typically, zero-shot learning approaches aim to recognize instances from the unseen or unknown testing *target* categories by transferring information, through intermediate-level semantic representations, from known observed *source* (or auxiliary) categories for which many labeled instances exist. In other words, supervised classes/instances, are used as context for recognition of classes that contain no visual instances at training time, but that can be put in some correspondence with supervised classes/instances. As such, a general experimental setting of ZSL is that the classes in target and source (auxiliary)

dataset are disjoint and typically the learning is done on the source dataset and then information is transferred to the target dataset, with performance measured on the latter.

This setting has a few important drawbacks: (1) it assumes that target classes cannot be mis-classified as source classes and vice versa; this greatly and unrealistically simplifies the problem; (2) the target label set is often relatively small, between ten [25] and several thousand unknown labels [14], compared to at least $30,000$ entry level categories that humans can distinguish; (3) large amounts of data in the source (auxiliary) classes are required, which is problematic as it has been shown that most object classes have only few instances (long-tailed distribution of objects in the world [40]); and (4) the vast open set vocabulary from semantic knowledge, defined as part of ZSL [31], is not leveraged in any way to inform the learning or source class recognition.

A few works recently looked at resolving (1) through class-incremental learning [38, 39] which is designed to distinguish between seen (source) and unseen (target) classes at the testing time and apply an appropriate model – supervised for the former and ZSL for the latter. However, (2)–(4) remain largely unresolved. In particular, while (2) and (3) are artifacts of the ZSL setting, (4) is more fundamental. For example, consider learning about a *car* by looking at image instances in Fig.1. Not knowing that other motor vehicles exist in the world, one may be tempted to call anything that has 4-wheels a *car*. As a result the zero-shot class *truck* may have large overlap with the *car* class (see Fig.1 [SVR]). However, imagine knowing that there also exist many other motor vehicles (trucks, mini-vans, *etc*). Even without having visually seen such objects, the very basic knowledge that they *exist* in the world and are closely related to a *car* should, in principal, alter the criterion for recognizing instance as a *car* (making the recognition criterion stricter in this case). Encoding this in our [SS-Voc] model results in better separation among classes.

To tackle the limitations of ZSL and towards the goal of generic open set recognition, we propose the idea of *semi-supervised vocabulary-informed learning*. Specifically, assuming we have few labeled training instances and a large open set vocabulary/semantic dictionary (along with textual sources from which statistical semantic relations among vocabulary atoms can be learned), the task of semi-supervised vocabulary-informed learning is to learn a model that utilizes semantic dictionary to help train better classifiers for observed (source) classes and unobserved (target) classes in supervised, zero-shot and open set image recognition settings. Different from standard semi-supervised learning, we do not assume unlabeled data is available, to help train classifier, and only *vocabulary* over the target classes is known.

Contributions: Our main contribution is to propose a novel paradigm for potentially open set image recognition: *semi-supervised vocabulary-informed learning* (**SS-Voc**), which is capable of utilizing vocabulary over unsupervised items, during training, to improve recognition. A unified maximum margin framework is used to encode this idea in practice. Particularly, classification is done through nearest-neighbor distance to class prototypes in the semantic embedding space, and we encode a set of constraints ensuring that labeled images project into semantic space such that they end up closer to the correct class prototypes than to incorrect ones (whether those prototypes are part of the source or target classes). We show that word embedding (word2vec) can be used effectively to initialize the semantic space. Experimentally, we illustrate that through this paradigm: we can achieve competitive supervised (on source classes) and ZSL (on target classes) performance, as well as open set image recognition performance with large number of unobserved vocabulary entities (up to $300,000$); effective learning with few samples is also illustrated.

2. Related Work

One-shot Learning: While most of machine learning-based object recognition algorithms require large amount of training data, one-shot learning [12] aims to learn object classifiers from one, or only few examples. To compensate for the lack of training instances and enable one-shot learning, *knowledge* much be transferred from other sources, for example, by sharing features [3], semantic attributes [17, 25, 34, 35], or contextual information [41]. However, none of previous works had used the open set vocabulary to help learn the object classifiers.

Zero-shot Learning: ZSL aims to recognize novel classes with no training instance by transferring *knowledge* from source classes. ZSL was first explored with use of attribute-based semantic representations [11, 15, 17, 18, 24, 32]. This required pre-defined attribute vector prototypes for each class, which is costly for a large-scale dataset. Recently, semantic word vectors were proposed as a way to embed any class name without human annotation effort; they can therefore serve as an alternative semantic representation [2, 14, 19, 30] for ZSL. Semantic word vectors are learned from large-scale text corpus by language models, such as word2vec [29], or GloVec [33]. However, most of previous work only use word vectors as semantic representations in ZSL setting, but have neither (1) utilized semantic word vectors explicitly for learning better classifiers; nor (2) for extending ZSL setting towards open set image recognition. A notable exception is [30] which aims to recognize 21K zero-shot classes given a modest vocabulary of 1K source classes; we explore vocabularies that are up to an order of the magnitude larger – 310K.

Open-set Recognition: The term "open set recognition" was initially defined in [37, 38] and formalized in [4, 36] which mainly aims at identifying whether an image belongs

to a seen or unseen classes. It is also known as class-incremental learning. However, none of them can further identify classes for unseen instances. An exception is [30] which augments zero-shot (unseen) class labels with source (seen) labels in some of their experimental settings. Similarly, we define the *open set image recognition* as the problems of recognizing the class name of an image from a potentially very large open set vocabulary (including, but not limited to source and target labels). Note that methods like [37, 38] are orthogonal but potentially useful here – it is still worth identifying seen or unseen instances to be recognized with different label sets as shown in experiments. Conceptually similar, but different in formulation and task, open-vocabulary object retrieval [20] focused on retrieving objects using natural language open-vocabulary queries.

Visual-semantic Embedding: Mapping between visual features and semantic entities has been explored in two ways: (1) directly learning the embedding by regressing from visual features to the semantic space using Support Vector Regressors (SVR) [11, 25] or neural network [39]; (2) projecting visual features and semantic entities into a common *new* space, such as SJE [2], WSABIE [44], ALE [1], DeViSE [14], and CCA [16, 18]. In contrast, our model trains a better visual-semantic embedding from only few training instances with the help of large amount of open set vocabulary items (using a maximum margin strategy). Our formulation is inspired by the unified semantic embedding model of [21], however, unlike [21], our formulation is built on word vector representation, contains a data term, and incorporates constraints to unlabeled vocabulary prototypes.

3. Vocabulary-informed Learning

Assume a labeled source dataset $\mathcal{D}_s = \{\mathbf{x}_i, z_i\}_{i=1}^{N_s}$ of N_s samples, where $\mathbf{x}_i \in \mathbb{R}^p$ is the image feature representation of image i and $z_i \in \mathcal{W}_s$ is a class label taken from a set of English words or phrases $\mathcal{W}$; consequently, $|\mathcal{W}_s|$ is the number of source classes. Further, suppose another set of class labels for target classes $\mathcal{W}_t$, such that $\mathcal{W}_s \cap \mathcal{W}_t = \emptyset$, for which no labeled samples are available. We note that potentially $|\mathcal{W}_t| >> |\mathcal{W}_s|$. Given a new test image feature vector $\mathbf{x}^*$ the goal is then to learn a function $z^* = f(\mathbf{x}^*)$, using all available information, that predicts a class label z^*. Note that the form of the problem changes drastically depending on which label set assumed for z^*: Supervised learning: $z^* \in \mathcal{W}_s$; Zero-shot learning: $z^* \in \mathcal{W}_t$; Open set recognition: $z^* \in \{\mathcal{W}_s, \mathcal{W}_t\}$ or, more generally, $z^* \in \mathcal{W}$. We posit that a single unified $f(\mathbf{x}^*)$ can be learned for all three cases. We formalize the definition of semi-supervised vocabulary-informed learning (SS-Voc) as follows:

Definition 3.1. *Semi-supervised Vocabulary-informed Learning (SS-Voc):* is a learning setting that makes use of complete vocabulary data ($\mathcal{W}$) during training. Unlike a more traditional ZSL that typically makes use of the vocabulary (*e.g.*, semantic embedding) at test time, SS-Voc utilizes exactly the same data during training. Notably, SS-Voc requires no additional annotations or semantic knowledge; it simply shifts the burden from testing to training, leveraging the vocabulary to learn a better model.

The vocabulary $\mathcal{W}$ can come from a semantic embedding space learned by word2vec [29] or GloVec [33] on large-scale corpus; each vocabulary entity $w \in \mathcal{W}$ is represented as a distributed semantic vector $\mathbf{u} \in \mathbb{R}^d$. Semantics of embedding space help with knowledge transfer among classes, and allow ZSL and open set image recognition. Note that such semantic embedding spaces are equivalent to the "semantic knowledge base" for ZSL defined in [31] and hence make it appropriate to use SS-Voc in ZSL setting.

Assuming we can learn a mapping $g : \mathbb{R}^p \rightarrow \mathbb{R}^d$, from image features to this semantic space, recognition can be carried out using simple nearest neighbor distance, *e.g.*, $f(\mathbf{x}^*) = car$ if $g(\mathbf{x}^*)$ is closer to $\mathbf{u}_{car}$ than to any other word vector; $\mathbf{u}_j$ in this context can be interpreted as the prototype of the class j. Thus the core question is then how to learn the mapping $g(\mathbf{x})$ and what form of inference is optimal in the semantic space. For learning we propose discriminative maximum margin criterion that ensures that labeled samples $\mathbf{x}_i$ project closer to their corresponding class prototypes $\mathbf{u}_{z_i}$ than to any other prototype $\mathbf{u}_i$ in the open set vocabulary $i \in \mathcal{W} \setminus z_i$.

3.1. Learning Embedding

Our maximum margin vocabulary-informed embedding learns the mapping $g(\mathbf{x}) : \mathbb{R}^p \rightarrow \mathbb{R}^d$, from low-level features $\mathbf{x}$ to the semantic word space by utilizing maximum margin strategy. Specifically, consider $g(\mathbf{x}) = W^T\mathbf{x}$, where[1] $W \subseteq \mathbb{R}^{p \times d}$. Ideally we want to estimate W such that $\mathbf{u}_{z_i} = W^T\mathbf{x}_i$ for all labeled instances in $\mathcal{D}_s$ (we would obviously want this to hold for instances belonging to unobserved classes as well, but we cannot enforce this explicitly in the optimization as we have no labeled samples for them).

Data Term: The easiest way to enforce the above objective is to minimize Euclidian distance between sample projections and appropriate prototypes in the embedding space[2]:

$$D\left(\mathbf{x}_i, \mathbf{u}_{z_i}\right) = \| W^T\mathbf{x}_i - \mathbf{u}_{z_i} \|_2^2 . \quad (1)$$

We need to minimize this term with respect to each instance $(\mathbf{x}_i, \mathbf{u}_{z_i})$, where z_i is the class label of instance $\mathbf{x}_i$ in $\mathcal{D}_s$. To prevent overfitting, we further regularize the solution:

$$\mathcal{L}\left(\mathbf{x}_i, \mathbf{u}_{z_i}\right) = D\left(\mathbf{x}_i, \mathbf{u}_{z_i}\right) + \lambda \| W \|_F^2, \quad (2)$$

where $\| \cdot \|_F$ indicates the Frobenius Norm. Solution to the Eq.(2) can be obtained through ridge regression.

[1] Generalizing to a kernel version is straightforward, see [43].

[2] Eq.(1) is also called data embedding [21] / compatibility function [2].

Nevertheless, to make the embedding more comparable to support vector regression (SVR), we employ the maximal margin strategy – $\epsilon-$insensitive smooth SVR ($\epsilon-$SSVR) [27] to replace the least square term in Eq.(2). That is,

$$\mathcal{L}\left(\mathbf{x}_i, \mathbf{u}_{z_i}\right) = \mathcal{L}_\epsilon\left(\mathbf{x}_i, \mathbf{u}_{z_i}\right) + \lambda \parallel W \parallel_F^2, \tag{3}$$

where $\mathcal{L}_\epsilon\left(\mathbf{x}_i, \mathbf{u}_{z_i}\right) = \mathbf{1}^T \mid \xi \mid_\epsilon^2$; λ is regularization coefficient. $\left(|\xi|_\epsilon\right)_j = \max\left\{0, |W_{\star j}^T \mathbf{x}_i - \left(\mathbf{u}_{z_i}\right)_j| - \epsilon\right\}$, $|\xi|_\epsilon \in \mathbb{R}^d$, and $()_j$ indicates the j-th value of corresponding vector. $W_{\star j}$ is the j-th column of W. The conventional $\epsilon-$SVR is formulated as a constrained minimization problem, *i.e.*, convex quadratic programming problem, while $\epsilon-$SSVR employs quadratic smoothing [47] to make Eq.(3) differentiable everywhere, and thus $\epsilon-$SSVR can be solved as an unconstrained minimization problem directly[3].

Pairwise Term: Data term above only ensures that labelled samples project close to their correct prototypes. However, since it is doing so for many samples and over a number of classes, it is unlikely that all the data constraints can be satisfied exactly. Specifically, consider the following case, if $\mathbf{u}_{z_i}$ is in the part of the semantic space where no other entities live (*i.e.*, distance from $\mathbf{u}_{z_i}$ to any other prototype in the embedding space is large), then projecting $\mathbf{x}_i$ further away from $\mathbf{u}_{z_i}$ is asymptomatic, *i.e.*, will not result in misclassification. However, if the $\mathbf{u}_{z_i}$ is close to other prototypes then minor error in regression may result in misclassification.

To embed this intuition into our learning, we enforce more discriminative constraints in the learned semantic embedding space. Specifically, the distance of $D\left(\mathbf{x}_i, \mathbf{u}_{z_i}\right)$ should not only be as close as possible, but should also be smaller than the distance $D\left(\mathbf{x}_i, \mathbf{u}_a\right)$, $\forall a \neq z_i$. Formally, we define the vocabulary pairwise maximal margin term[4]

$$\mathcal{M}_V\left(\mathbf{x}_i, \mathbf{u}_{z_i}\right) = \frac{1}{2}\sum_{a=1}^{A_V}\left[C + \frac{1}{2}D\left(\mathbf{x}_i, \mathbf{u}_{z_i}\right) - \frac{1}{2}D\left(\mathbf{x}_i, \mathbf{u}_a\right)\right]_+^2 \tag{4}$$

where $a \in \mathcal{W}_t$ is selected from the open vocabulary; C is the margin gap constant. Here, $[\cdot]_+^2$ indicates the quadratically smooth hinge loss [47] which is convex and has the gradient at every point. To speedup computation, we use the closest A_V target prototypes to each source/auxiliary prototype $\mathbf{u}_{z_i}$ in the semantic space. We also define similar constraints for the source prototype pairs:

$$\mathcal{M}_S\left(\mathbf{x}_i, \mathbf{u}_{z_i}\right) = \frac{1}{2}\sum_{b=1}^{B_S}\left[C + \frac{1}{2}D\left(\mathbf{x}_i, \mathbf{u}_{z_i}\right) - \frac{1}{2}D\left(\mathbf{x}_i, \mathbf{u}_b\right)\right]_+^2 \tag{5}$$

[3]We found Eq.(2) and Eq.(3) have similar results, on average, but formulation in Eq.(3) is more stable and has lower variance.

[4]Crammer and Singer loss [42, 8] is the upper bound of Eq (4) and (5) which we use to tolerate variants of $\mathbf{u}_{z_i}$ (e.g. 'pigs' Vs. 'pig' in Fig. 2) and thus are better for our tasks.

where $b \in \mathcal{W}_s$ is selected from source/auxiliary dataset vocabulary. This term enforces that $D\left(\mathbf{x}_i, \mathbf{u}_{z_i}\right)$ should be smaller than the distance $D\left(\mathbf{x}_i, \mathbf{u}_b\right)$, $\forall b \neq z_i$. To facilitate the computation, we similarly use closest B_S prototypes that are closest to each prototype $\mathbf{u}_{z_i}$ in the source classes. Our complete pairwise maximum margin term is:

$$\mathcal{M}\left(\mathbf{x}_i, \mathbf{u}_{z_i}\right) = \mathcal{M}_V\left(\mathbf{x}_i, \mathbf{u}_{z_i}\right) + \mathcal{M}_S\left(\mathbf{x}_i, \mathbf{u}_{z_i}\right). \tag{6}$$

We note that the form of rank hinge loss in Eq.(4) and Eq.(5) is similar to DeViSE [14], but DeViSE only considers loss with respect to source/auxiliary data and prototypes.

Vocabulary-informed Embedding: The complete combined objective can now be written as:

$$W = \underset{W}{\operatorname{argmin}} \sum_{i=1}^{n_T}(\alpha\mathcal{L}_\epsilon\left(\mathbf{x}_i, \mathbf{u}_{y_i}\right) + (1-\alpha)\mathcal{M}\left(\mathbf{x}_i, \mathbf{u}_{z_i}\right)) + \lambda \parallel W \parallel_F^2, \tag{7}$$

where $\alpha \in [0, 1]$ is ratio coefficient of two terms. One practical advantage is that the objective function in Eq.(7) is an unconstrained minimization problem which is differentiable and can be solved with L-BFGS. W is initialized with all zeros and converges in $10-20$ iterations.

Fine-tuning Word Vector Space: Above formulation works well assuming semantic space is well laid out and linear mapping is sufficient. However, we posit that word vector space itself is not necessarily optimal for visual discrimination. Consider the following case: two visually similar categories may appear far away in the semantic space. In such a case, it would be difficult to learn a linear mapping that matches instances with category prototypes properly. Inspired by this intuition, which has also been expressed in natural language models [6], we propose to fine-tune the word vector representation for better visual discriminability.

One can potentially fine-tune the representation by optimizing $\mathbf{u}_i$ directly, in an alternating optimization (*e.g.*, as in [21]). However, this is only possible for source/auxiliary class prototypes and would break regularities in the semantic space, reducing ability to transfer knowledge from source/auxilary to target classes. Alternatively, we propose optimizing a global warping, V, on the word vector space:

$$\{W, V\} = \underset{W,V}{\operatorname{argmin}} \sum_{i=1}^{n_T}(\alpha\mathcal{L}_\epsilon\left(\mathbf{x}_i, \mathbf{u}_{y_i}V\right) + (1-\alpha)\,\mathcal{M}\left(\mathbf{x}_i, \mathbf{u}_{z_i}V\right)) + \lambda \parallel W \parallel_F^2 + \mu \parallel V \parallel_F^2, \tag{8}$$

where μ is regularization coefficient. Eq.(8) can still be solved using L-BFGS and V is initialized using an identity matrix. The algorithm first updates W and then V; typically the step of updating V can converge within 10 iterations and the corresponding class prototypes used for final classification are updated to be $\mathbf{u}_{z_i}V$.

3.2. Maximum Margin Embedding Recognition

Once embedding model is learned, recognition in the semantic space can be done in a variety of ways. We explore a simple alternative to classify the testing instance $\mathbf{x}^\star$,

$$z^* = \underset{i}{\operatorname{argmin}} \parallel W\mathbf{x}^* - \phi\left(\mathbf{u}_i, V, W, \mathbf{x}^*\right) \parallel_2^2 . \quad (9)$$

Nearest Neighbor (NN) classifier directly measures the distance between predicted semantic vectors with the prototypes in semantic space, *i.e.*, $\phi\left(\mathbf{u}_i, V, W, \mathbf{x}^*\right) = \mathbf{u}_i V$. We further employ the k-nearest neighbors (KNN) of testing instances to average the predictions, *i.e.*, $\phi\left(\cdot\right)$ is averaging the KNN instances of predicted semantic vectors.[5]

4. Experiments

Datasets. We conduct our experiments on Animals with Attributes (AwA) dataset, and ImageNet 2012/2010 dataset. AwA consists of 50 classes of animals (30,475 images in total). In [25] standard split into 40 source/auxiliary classes ($|\mathcal{W}_s| = 40$) and 10 target/test classes ($|\mathcal{W}_t| = 10$) is introduced. We follow this split for supervised and zero-shot learning. We use OverFeat features (downloaded from [19]) on AwA to make the results more easily comparable to state-of-the-art. ImageNet 2012/2010 dataset is a large-scale dataset. We use 1000 ($|\mathcal{W}_s| = 1000$) classes of ILSVRC 2012 as the source/auxiliary classes and 360 ($|\mathcal{W}_t| = 360$) classes of ILSVRC 2010 that are not used in ILSVRC 2012 as target data. We use pre-trained VGG-19 model [7] to extract deep features for ImageNet. On both dataset, we use few instances from source dataset to mimic human performance of learning from few examples and ability to generalize.

Recognition tasks. We consider three different settings in a variety of experiments (in each experiment we carefully denote which setting is used):

SUPERVISED recognition, where learning is on source classes and we assume test instances come from same classes with $\mathcal{W}_s$ as recognition vocabulary;

ZERO-SHOT recognition, where learning is on source classes and we assume test instances coming from target dataset with $\mathcal{W}_t$ as recognition vocabulary;

OPEN-SET recognition, where we use entirely open vocabulary with $|\mathcal{W}| \approx 310K$ and use test images from both source and target splits.

Competitors. We compare the following models,

SVM: SVM classifier trained directly on the training instances of source data, without the use of semantic embedding. This is the standard (SUPERVISED) learning setting and the learned classifier can only predict the labels in testing data of source classes.

SVR-Map: SVR is used to learn W and the recognition is done in the resulting semantic manifold. This corresponds to only using Eq.(3) to learn W.

DeVise, ConSE, AMP: To compare with state-of-the-art large-scale zero-shot learning approaches we implement DeViSE [14] and ConSE [30].[6] ConSE uses a multi-class logistic regression classifier for predicting class probabilities of source instances; and the parameter T (number of top-T nearest embeddings for a given instance) was selected from $\{1, 10, 100, 1000\}$ that gives the best results. ConSE method in supervised setting works the same as SVR. We use the AMP code provided on the author webpage [19].

SS-Voc: We test three different variants of our method.

closed is a variant of our maximum margin leaning of W with the vocabulary-informed constraints only from known classes (*i.e.*, closed set $\mathcal{W}_s$).

W corresponds to our full model with maximum margin constraints coming from both $\mathcal{W}_s$ and $\mathcal{W}_t$ (or $\mathcal{W}$). We compute W using Eq.(7), but without optimizing V.

full further fine-tunes the word vector space by also optimizing V using Eq.(8).

Open set vocabulary. We use google word2vec to learn the open set vocabulary set from a large text corpus of around 7 billion words: UMBC WebBase (3 billion words), the latest Wikipedia articles (3 billion words) and other web documents (1 billion words). Some rare (low frequency) words and high frequency stopping words were pruned in the vocabulary set: we remove words with the frequency < 300 or $> 10\,million$ times. The result is a vocabulary of around 310K words/phrases with $openness \approx 1$, which is defined as $openness = 1 - \sqrt{(2 \times |\mathcal{W}_s|) / (|\mathcal{W}|)}$. [38].

Computational and parameters selection and scalability. All experiments are repeated 10 times, to avoid noise due to small training set size, and we report an average across all runs. For all the experiments, the mean accuracy is reported, *i.e.*, the mean of the diagonal of the confusion matrix on the prediction of testing data. We fix the parameters μ and λ as 0.01 and $\alpha = 0.6$ in our experiments when only few training instances are available for AwA (5 instances per class) and ImageNet (3 instances per class). Varying values of λ, μ and α leads to $< 1\%$ variances on AwA and $< 0.2\%$ variances on ImageNet dataset; but the experimental conclusions still hold. Cross-validation is conducted when

[5] This strategy is known as Rocchio algorithm in information retrieval. Rocchio algorithm is a method for relevance feedback by using more relevant instances to update the query instances for better recall and possibly precision in vector space (Chap 14 in [28]). It was first suggested for use on ZSL in [17]; more sophisticated algorithms [16, 34] are also possible.

[6] Code for [14] and [30] is not publicly available.

	Testing Classes							SS-Voc		
	Aux	Targ.	Total	Vocab	Chance	SVM	SVR	closed	W	full
SUPERVISED	✓		40	40	2.5	52.1	51.4/57.1	52.9/58.2	53.6/58.6	53.9/59.1
ZERO-SHOT		✓	10	10	10	-	52.1/58.0	58.6/60.3	59.5/68.4	61.1/68.9

Table 1. Classification accuracy (%) on AwA dataset for SUPERVISED and ZERO-SHOT settings for 100/1000-dim word2vec representation.

more training instances are available. A_V and B_S are set to 5 to balance computational cost and efficiency of pairwise constraints.

To solve Eq.(8) at a scale, one can use Stochastic Gradient Descent (SGD) which makes great progress initially, but often is slow when approaches a solution. In contrast, the L-BFGS method mentioned above can achieve steady convergence at the cost of computing the full objective and gradient at each iteration. L-BFGS can usually achieve better results than SGD with good initialization, however, is computationally expensive. To leverage benefits of both of these methods, we utilize a hybrid method to solve Eq.(8) in large-scale datasets: the solver is initialized with few instances to approximate the gradients using SGD first, then gradually more instances are used and switch to L-BFGS is made with iterations. This solver is motivated by Friedlander *et al.* [13], who theoretically analyzed and proved the convergence for the hybrid optimization methods. In practice, we use L-BFGS and the Hybrid algorithms for AwA and ImageNet respectively. The hybrid algorithm can save between $20 \sim 50\%$ training time as compared with L-BFGS.

4.1. Experimental results on AwA dataset

We report AwA experimental results in Tab. 1, which uses 100/1000-dimensional word2vec representation (*i.e.*, $d = 100/1000$). We highlight the following observations: (1) **SS-Voc** variants have better classification accuracy than SVM and SVR. This validates the effectiveness of our model. Particularly, the results of our **SS-Voc:full** are $1.8/2\%$ and $9/10.9\%$ higher than those of SVR/SVM on supervised and zero-shot recognition respectively. Note that though the results of SVM/SVR are good for supervised recognition tasks (52.1 and 51.4/57.1 respectively), we can further improve them, which we attribute to the more discriminative classification boundary informed by the vocabulary. (2) **SS-Voc:**W significantly, by up to 8.1%, improves zero-shot recognition results of **SS-Voc:closed**. This validates the importance of information from open vocabulary. (3) **SS-Voc** benefits more from open set vocabulary as compared to word vector space fine-tuneing. The results of supervised and zero-shot recognition of **SS-Voc:full** are $1/0.9\%$ and $2.5/8.6\%$ higher than those of **SS-Voc:closed**.

Comparing to state-of-the-art on ZSL: We compare our results with the state-of-the-art ZSL results on AwA dataset in Tab. 2. We compare **SS-Voc:full** trained with all source instances, 800 (20 instances / class), and 200 instances (5 instances / class). Our model achieves 78.3% accuracy, which is remarkably higher than all previous methods. This is particularly impressive taking into account the fact that we use only a semantic space and no additional attribute representations that many other competitor methods utilize. Further, our results with 800 training instances, a small fraction of the $24,295$ instances used to train all other methods, already outperform all other approaches. We argue that much of our success and improvement comes from a more discriminative information obtained using an open set vocabulary and corresponding large margin constraints, rather than from the features, since our method improved 25.1% as compared with DAP [25] which uses the same OverFeat features. Note, our **SS-Voc:full** result is 4.4% higher than the closest competitor [2]; this improvement is statistically significant. Comparing with our work, [2] did not only use more powerful visual features (GoogLeNet Vs. OverFeat), but also employed more semantic embeddings (attributes, GloVe[7] and WordNet-derived similarity embeddings as compared to our word2vec).

Methods	S. Sp	Features	Acc.
SS-Voc:full	W	CNN_{OverFeat}	**78.3**
800 instances	W	CNN_{OverFeat}	**74.4**
200 instances	W	CNN_{OverFeat}	68.9
Akata *et al.* [2]	A+W	$CNN_{\text{GoogleLeNet}}$	73.9
TMV-BLP [16]	A+W	CNN_{OverFeat}	69.9
AMP (SR+SE) [19]	A+W	CNN_{OverFeat}	66.0
DAP [25]	A	CNN_{VGG19}	57.5
PST[34]	A+W	CNN_{OverFeat}	54.1
DAP [25]	A	CNN_{OverFeat}	53.2
DS [35]	W/A	CNN_{OverFeat}	52.7
Jayaraman *et al.* [22]	A	low-level	48.7
Yu *et al.* [46]	A	low-level	48.3
IAP [25]	A	CNN_{OverFeat}	44.5
HEX [9]	A	CNN_{DECAF}	44.2
AHLE [1]	A	low-level	43.5

Table 2. **Zero-shot comparison on AwA.** We compare the state-of-the-art ZSL results using different semantic spaces (S. Sp) including word vector (W) and attribute (A). 1000 dimension word2vec dictionary is used for SS-Voc. (Chance-level =10%). Different types of CNN and hand-crafted low-level feature are used by different methods. Except SS-Voc (200/800), all instances of source data (24295 images) are used for training. As a general reference, the classification accuracy on ImageNet: $CNN_{\text{DECAF}} < CNN_{\text{OverFeat}} < CNN_{\text{VGG19}} < CNN_{\text{GoogleLeNet}}$.

[7] GloVe [33] can be taken as an improved version of word2vec.

Large-scale open set recognition: Here we focus on OPEN-SET$_{310K}$ setting with the large vocabulary of approximately 310K entities; as such the chance performance of the task is much much lower. In addition, to study the effect of performance as a function of the open vocabulary set, we also conduct two additional experiments with different label sets: (1) OPEN-SET$_{1K-NN}$: the 1000 labels from nearest neighbor set of ground-truth class prototypes are selected from the complete dictionary of 310K labels. This corresponds to an open set fine grained recognition; (2) OPEN-SET$_{1K-RND}$: 1000 label names randomly sampled from 310K set. The results are shown in Fig. 2. Also note that we did not fine-tune the word vector space (*i.e.*, V is an Identity matrix) on OPEN-SET$_{310K}$ setting since Eq (8) can optimize a better visual discriminability only on a relative small subset as compared with the 310K vocabulary. While our OPEN-SET variants do not assume that test data comes from either source/auxiliary domain or target domain, we split the two cases to mimic SUPERVISED and ZERO-SHOT scenarios for easier analysis.

On SUPERVISED-like setting, Fig. 2 (left), our accuracy is better than that of SVR-Map on all the three different label sets and at all hit rates. The better results are largely due to the better embedding matrix W learned by enforcing maximum margins between training class name and open set vocabulary on source training data.

On ZERO SHOT-like setting, our method still has a notable advantage over that of SVR-Map method on Top-k ($k > 5$) accuracy, again thanks to the better embedding W learned by Eq. (7). However, we notice that our top-1 accuracy on ZERO SHOT-like setting is lower than SVR-Map method. We find that our method tends to label some instances from target data with their nearest classes from within source label set. For example, "humpback whale" from testing data is more likely to be labeled as "blue whale". However, when considering Top-k ($k > 5$) accuracy, our method still has advantages over baselines.

4.2. Experimental results on ImageNet dataset

We further validate our findings on large-scale ImageNet 2012/2010 dataset; 1000-dimensional word2vec representation is used here since this dataset has larger number of classes than AwA. We highlight that our results are still better than those of two baselines – SVR-Map and SVM on (SUPERVISED) and (ZERO-SHOT) settings respectively as shown in Tab. 3. The open set image recognition results are shown in Fig. 4. On both SUPERVISED-like and ZERO-SHOT-like settings, clearly our framework still has advantages over the baseline which directly matches the nearest neighbors from the vocabulary by using predicted semantic word vectors of each testing instance.

We note that SUPERVISED SVM results (34.61%) on ImageNet are lower than 63.30% reported in [7], despite using the same features. This is because only few, 3 samples per class, are used to train our models to mimic human performance of learning from few examples and illustrate ability of our model to learn with little data. However, our semi-supervised vocabulary-informed learning can improve the recognition accuracy on all settings. On open set image recognition, the performance has dropped from 37.12% (SUPERVISED) and 8.92% (ZERO-SHOT) to around 9% and 1% respectively (Fig. 4). This drop is caused by the intrinsic difficulty of the open set image recognition task ($\approx 300\times$ increase in vocabulary) on a large-scale dataset. However, our performance is still better than the SVR-Map baseline which in turn significantly better than the chance-level.

AwA Dataset	Testing Classes Aux.	Targ.	Total	Vocab
OPEN-SET$_{1K-NN}$	(left)	(right)	40 / 10	1000*
OPEN-SET$_{1K-RND}$			40 / 10	1000†
OPEN-SET$_{310K}$			40 / 10	310K

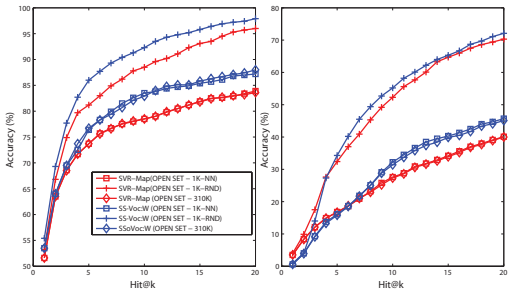

Figure 2. **Open set recognition results on AwA dataset:** Openness=0.9839. Chance=$3.2e-4\%$. Ground truth label is extended for its variants. For example, we count a correct label if a 'pig' image is labeled as 'pigs'. $\star$,†:different 1000 label settings.

We also evaluated our model with larger number of training instances (> 3 per class). We observe that for standard supervised learning setting, the improvements achieved using vocabulary-informed learning tend to somewhat diminish as the number of training instances substantially grows. With large number of training instances, the mapping between low-level image features and semantic words, $g(\mathbf{x})$, becomes better behaved and effect of additional constraints, due to the open-vocabulary, becomes less pronounced.

Comparing to state-of-the-art on ZSL. We compare our results to several state-of-the-art large-scale zero-shot recognition models. Our results, **SS-Voc:full**, are better than those of ConSE, DeViSE and AMP on both T-1 and T-5 metrics with a *very* significant margin (improvement over best competitor, ConSE, is 3.43 percentage points or nearly 62% with $3,000$ training samples). Poor results of DeViSE with $3,000$ training instances are largely due to the inefficient learning of visual-semantic embedding matrix. AMP algorithm also relies on the embedding matrix from DeViSE, which explains similar poor performance of AMP

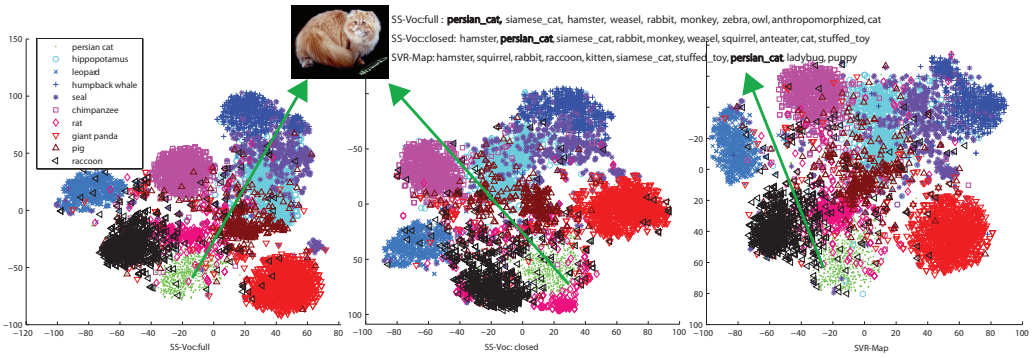

Figure 3. t-SNE visualization of AwA 10 testing classes. Please refer to Supplementary material for larger figure.

	Testing Classes							**SS-Voc**		
	Aux	Targ.	Total	Vocab	Chance	SVM	SVR	closed	W	full
SUPERVISED	✓		1000	1000	0.1	33.8	25.6	34.2	36.3	37.1
ZERO-SHOT		✓	360	360	0.278	-	4.1	8.0	8.2	8.9

Table 3. The classification accuracy (%) of ImageNet 2012/2010 dataset on SUPERVISED and ZERO-SHOT settings.

	Testing Classes			
ImageNet Data	Aux.	Targ.	Total	Vocab
OPEN-SET$_{310K}$	(left)	(right)	1000 / 360	$310K$

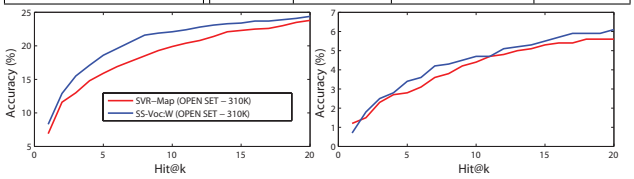

Figure 4. **Open set recognition results on ImageNet 2012/2010 dataset:** Openness=0.9839. Chance=$3.2e-4\%$. We use the synsets of each class— a set of synonymous (word or phrase) terms as the ground truth names for each instance.

with $3,000$ training instances. In contrast, our **SS-Voc:full** can leverage discriminative information from open vocabulary and max-margin constraints, which helps improve performance. For DeViSE with *all* ImageNet instances, we confirm the observation in [30] that results of ConSE are much better than those of DeViSE. Our results are a further significant improved from ConSE.

4.3. Qualitative results of open set image recognition

t-SNE visualization of AwA 10 target testing classes is shown in Fig. 3. We compare our **SS-Voc:full** with **SS-Voc:closed** and SVR. We note that (1) the distributions of 10 classes obtained using **SS-Voc** are more centered and more separable than those of SVR (*e.g.*, *rat*, *persian cat* and *pig*), due to the data and pairwise maximum margin terms that help improve the generalization of $g(\mathbf{x})$ learned; (2) the distribution of different classes obtained using the full model **SS-Voc:full** are also more separable than those of **SS-Voc:closed**, *e.g.*, *rat*, *persian cat* and *raccoon*. This can be attributed to the addition of the open-vocabulary-informed constraints during learning of $g(\mathbf{x})$, which further improves generalization. For example, we show an open set recognition example image of "persian_cat", which is wrongly classified as a "hamster" by **SS-Voc:closed**.

Methods	S. Sp	Feat.	T-1	T-5
SS-Voc:full	W	CNN_{OverFeat}	**8.9/9.5**	**14.9/16.8**
ConSE [30]	W	CNN_{OverFeat}	5.5/7.8	13.1/15.5
DeViSE [14]	W	CNN_{OverFeat}	3.7/5.2	11.8/12.8
AMP [19]	W	CNN_{OverFeat}	3.5/6.1	10.5/13.1
Chance	–	–	$2.78e$-3	–

Table 4. **ImageNet comparison to state-of-the-art on ZSL:** We compare the results of using $3,000$/*all* training instances for all methods; T-1 (top 1) and T-5 (top 5) classification in % is reported.

Partial illustration of the embeddings learned for the ImageNet2012/2010 dataset are illustrated in Figure 1, where 4 source/auxiliary and 2 target/zero-shot classes are shown. Again better separation among classes is largely attributed to open-set max-margin constraints introduced in our **SS-Voc:full** model. Additional examples of miss-classified instances are available in the supplemental material.

5. Conclusion and Future Work

This paper introduces the problem of semi-supervised vocabulary-informed learning, by utilizing open set semantic vocabulary to help train better classifiers for observed and unobserved classes in supervised learning, ZSL and open set image recognition settings. We formulate semi-supervised vocabulary-informed learning in the maximum margin framework. Extensive experimental results illustrate the efficacy of such learning paradigm. Strikingly, it achieves competitive performance with only few training instances and is relatively robust to large open set vocabulary of up to $310,000$ class labels.

We rely on word2vec to transfer information between observed and unobserved classes. In future, other linguistic or visual semantic embeddings could be explored instead, or in combination, as part of vocabulary-informed learning.

References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *CVPR*, 2013.

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

[3] E. Bart and S. Ullman. Cross-generalization: learning novel classes from a single example by feature replacement. In *CVPR*, 2005.

[4] A. Bendale and T. Boult. Towards open world recognition. In *CVPR*, 2015.

[5] I. Biederman. Recognition by components - a theory of human image understanding. *Psychological Review*, 1987.

[6] S. R. Bowman, C. Potts, and C. D. Manning. Learning distributed word representations for natural logic reasoning. *CoRR*, abs/1410.4176, 2014.

[7] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*, 2014.

[8] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *JMLR*, 2001.

[9] J. Deng, N. Ding, Y. Jia, A. Frome, K. Murphy, S. Bengio, Y. Li, H. Neven, and H. Adam. Large-scale object classification using label relation graphs. In *ECCV*, 2014.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

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

[12] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE TPAMI*, 2006.

[13] M. P. Friedlander and M. Schmidt. Hybrid deterministic-stochastic methods for data fitting. *SIAM J. Scientific Computing*, 2012.

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

[15] Y. Fu, T. Hospedales, T. Xiang, and S. Gong. Attribute learning for understanding unstructured social activity. In *ECCV*, 2012.

[16] Y. Fu, T. M. Hospedales, T. Xiang, Z. Fu, and S. Gong. Transductive multi-view embedding for zero-shot recognition and annotation. In *ECCV*, 2014.

[17] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Learning multi-modal latent attributes. *IEEE TPAMI*, 2013.

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

[19] Z. Fu, T. Xiang, E. Kodirov, and S. Gong. zero-shot object recognition by semantic manifold distance. In *CVPR*, 2015.

[20] S. Guadarrama, E. Rodner, K. Saenko, N. Zhang, R. Farrell, J. Donahue, and T. Darrell. Open-vocabulary object retrieval. In *Robotics Science and Systems (RSS)*, 2014.

[21] S. J. Hwang and L. Sigal. A unified semantic embedding: relating taxonomies and attributes. In *NIPS*, 2014.

[22] D. Jayaraman and K. Grauman. Zero shot recognition with unreliable attributes. In *NIPS*, 2014.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[24] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009.

[25] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE TPAMI*, 2013.

[26] H. Larochelle, D. Erhan, and Y. Bengio. Zero-data learning of new tasks. In *AAAI*, 2008.

[27] Y.-J. Lee, W.-F. Hsieh, and C.-M. Huang. ϵ-SSVR: A smooth support vector machine for ϵ-insensitive regression. *IEEE TKDE*, 2005.

[28] C. D. Manning, P. Raghavan, and H. Schutze. *Introduction to Information Retrieval*. Cambridge University Press, 2009.

[29] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Neural Information Processing Systems*, 2013.

[30] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. *ICLR*, 2014.

[31] M. Palatucci, G. Hinton, D. Pomerleau, and T. M. Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, 2009.

[32] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011.

[33] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[34] M. Rohrbach, S. Ebert, and B. Schiele. Transfer learning in a transductive setting. In *NIPS*, 2013.

[35] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele. What helps where – and why? semantic relatedness for knowledge transfer. In *CVPR*, 2010.

[36] H. Sattar, S. Muller, M. Fritz, and A. Bulling. Prediction of search targets from fixations in open-world settings. In *CVPR*, 2015.

[37] W. J. Scheirer, L. P. Jain, and T. E. Boult. Probability models for open set recognition. *IEEE TPAMI*, 2014.

[38] W. J. Scheirer, A. Rocha, A. Sapkota, and T. E. Boult. Towards open set recognition. *IEEE TPAMI*, 2013.

[39] R. Socher, M. Ganjoo, H. Sridhar, O. Bastani, C. D. Manning, and A. Y. Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, 2013.

[40] A. Torralba, R. Fergus, and W. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE TPAMI*, 2008.

[41] A. Torralba, K. P. Murphy, and W. T. Freeman. Using the forest to see the trees: Exploiting context for visual object detection and localization. *Commun. ACM*, 2010.

[42] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *JMLR*, 2005.

[43] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. In *IEEE TPAMI*, 2011.

[44] J. Weston, S. Bengio, and N. Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI*, 2011.

[45] Z. Wu, Y. Fu, Y.-G. Jiang, and L. Sigal. Harnessing object and scene semantics for large-scale video understanding. In *CVPR*, 2016.

[46] F. X. Yu, L. Cao, R. S. Feris, J. R. Smith, and S.-F. Chang. Designing category-level attributes for discriminative visual recognition. *CVPR*, 2013.

[47] T. Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *ICML*, 2004.