# Multi-Label Zero-Shot Learning with Structured Knowledge Graphs

Chung-Wei Lee[1]*, Wei Fang[1]*, Chih-Kuan Yeh[2], Yu-Chiang Frank Wang[1]
Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan[1]
Machine Learning Department, Carnegie Mellon University, Pittsburgh, USA[2]
{b02901088,b02901054,ycwang}@ntu.edu.tw, cjyeh@cs.cmu.edu

## Abstract

*In this paper, we propose a novel deep learning architecture for multi-label zero-shot learning (ML-ZSL), which is able to predict multiple unseen class labels for each input instance. Inspired by the way humans utilize semantic knowledge between objects of interests, we propose a framework that incorporates knowledge graphs for describing the relationships between multiple labels. Our model learns an information propagation mechanism from the semantic label space, which can be applied to model the interdependencies between seen and unseen class labels. With such investigation of structured knowledge graphs for visual reasoning, we show that our model can be applied for solving multi-label classification and ML-ZSL tasks. Compared to state-of-the-art approaches, comparable or improved performances can be achieved by our method.*

## 1. Introduction

Real-world machine learning applications such as image annotation, music categorization, or medical diagnosis require assigning more than one class label to each input instance. Take image annotation for example, the learning models have to predict multiple labels like sky, sea, or ship for a single input image. Different from traditional multi-class methods which only predict one class label for each instance, learning multi-label classification models typically require additional efforts. More specifically, we not only need to relate the images with their multiple labels, it is often desirable to exploit label correlation due to the co-occurrences of the labels of interest.

In general, binary relevance [44] is the simplest solution to multi-label classification problems, which coverts the original task to multiple disjoint binary classification problems. However, it lacks the ability to model label co-occurrences, and thus might not be preferable. Approaches such as [38, 7] take cross-label correlation by assuming label priors, while label-embedding based methods [3, 43, 6, 5, 4] project both input images and their labels onto a latent space to exploit label correlation. Methods that utilize deep neural networks have also been proposed. BP-MLL [50] first proposed a loss function for modeling the dependency across labels, while other recent works proposed different loss functions [18, 34] or architectures [46, 45, 49] to further improve performance.

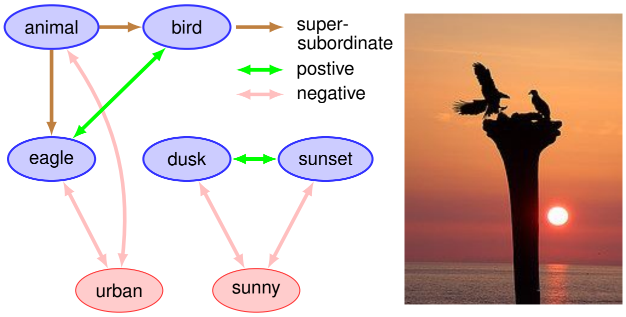

Figure 1. Illustration of structured knowledge graph for modeling the dependency between labels in the semantic space. We learn and utilize such graphs for relating the belief for each label, so that prediction of multiple seen or unseen labels can be achieved. The ground truth labels are noted in blue.

Extending from multi-label classification, multi-label zero-shot learning (ML-ZSL) is a branch of zero-shot learning (ZSL), which require the prediction of unseen labels which are not defined during training. Traditional multi-label approaches such as binary relevance or label-prior based methods obviously cannot be directly applied to ML-ZSL, since such methods lack the ability to generalize to unseen class labels. In contrast, approaches that utilize label representations in the semantic space such as label-embedding methods can be more easily adapted to ML-ZSL, given label representations of the unseen classes. Generally, label representations are obtained from human-annotated attribute vectors that describe the labels of interest either in a specific domain, or via distributed word embeddings learned from linguistic resources.

Nevertheless, although recent ML-ZSL methods such as

[*] Indicates equal contribution

[31, 16, 51, 17, 39] have been proposed, existing approaches typically do not take advantages of structured knowledge and reasoning. Humans recognize objects not only by appearance, but also by using knowledge of the world learned through experience. Inspired by the above observation, we focus on leveraging existing structural knowledge for ML-ZSL, with the goal of deriving proper dependencies between different label concepts for both seen and unseen ones. Figure 1 illustrates how knowledge graphs can help in this problem, where we can model the co-occurring and non-co-occurring concepts and extend this knowledge to unseen classes with an external structured knowledge graph. There has been work on multi-label problems utilizing structured knowledge. [10] introduced a graph representation that enforces certain relations between label concepts. [21] employed recurrent neural networks (RNN) [20, 41] to model positive and negative correlations between different concept layers. More recently, [30] extended neural networks for graphs [40, 27] to efficiently learn a model that reasons about different types of relationships between class labels by propagating information in a knowledge graph.

However, to the best of our knowledge, none of existing work advances structured knowledge reasoning for ML-ZSL. In this paper, we propose a novel ML-ZSL approach to observe and incorporate associated structured knowledge. Labels are represented with semantic vectors and an information propagation mechanism is learned from the label relations observed in the semantic space. The propagation of such label relation information is then used to modify the initial beliefs for each class label. Once the propagation process is complete, multi-label classification (or ML-ZSL) can be performed accordingly. Our model incorporates structured knowledge graphs observed from WordNet [33] into an end-to-end learning framework, while learning the label representations and information to be propagated in the semantic space. With this framework, we are able to achieve ZSL by assigning the unseen label embedding vector into our learning model. We will show the effectiveness of our model in advancing the structured knowledge for reasoning, which would benefit the task of ML-ZSL.

The main contributions of this work are highlighted as follows:

- To the best of our knowledge, our model is among the first to advance structured information and knowledge graphs for ML-ZSL.
- Our method advances a label propagation mechanism in the semantic space, enabling the reasoning of the learned model for predicting unseen labels.
- With comparable performance on standard multi-label classification tasks, our method performs favorably against recent models for ML-ZSL.

## 2. Related Work

Remarkable developments on image classification has been observed over the past few years due to the availability of large-scale datasets like ImageNet [11] and the development of deep convolutional neural networks [23, 19].

Among image classification tasks, multi-label classification aims at predicting multiple labels for an input image, whcih can be achieved by the technique of binary relevance [44] using neural networks. To further improve the performance, label co-occurrence and relations between labels are considered in recent works. Label embedding methods are among the popular techniques, which transform labels into embedded label vectors, so that the correlation between labels can be exploited [47, 43, 18, 29, 49]. As non-linear embedding approaches, deep neural networks have also been utilized for multi-label classification [50, 34, 18, 46, 45, 49].

Another way to determine the dependency between labels is via exploring explicit semantic relations between the labels. The Hierarchy and Exclusion (HEX) graph [10] captures semantic relations: mutual exclusion, overlap and subsumption between any two labels, improving object classification by exploiting the label relations. The model is further extended to allow for soft or probabilistic relations between labels [13]. Later, [21] introduced Structured Inference Neural Network (SINN). Inspired by the idea of Recurrent Neural Network (RNN) [20, 41], positive correlation and negative correlation between labels are derived for bidirectionally propagating information between concept layers, which further improves the classification performance; Focusing on single-label activity recognition, [12] view both activity of input image and actions of each person in that image as a graph, and utilize RNN to update the observed graph for activity prediction. On the other hand, Graph Neural Networks [40], [27] present architectures of Graph Gated Neural Networks (GGNN), which apply Gated Recurrent Units (GRU) [8] and allow propagation on the graphs. As a modification of GGNN, Graph Search Neural Network (GSNN) [30] is successfully applied for multi-label image classification to exploit explicit semantic relations in the form of structured knowledge graphs.

Different from multi-label classification, zero-shot learning (ZSL) is a challenging task, which needs to recognize test inputs as unseen categories. ZSL also attracts extensive attention from the vision community [36, 1, 42, 15, 22, 35, 16, 25, 2, 26], which is typically addressed by relating semantic information like attributes [24, 14] and word vectors [32, 37] to the presence of visual content.

Extended from ZSL, multi-label zero-shot learning (ML-ZSL) further requires one to assign multiple unseen labels for each instance. To solve ML-ZSL tasks, COSTA[31] assumes co-occurrence statistics and estimates classifiers for seen labels by weighted combinations of seen classes.

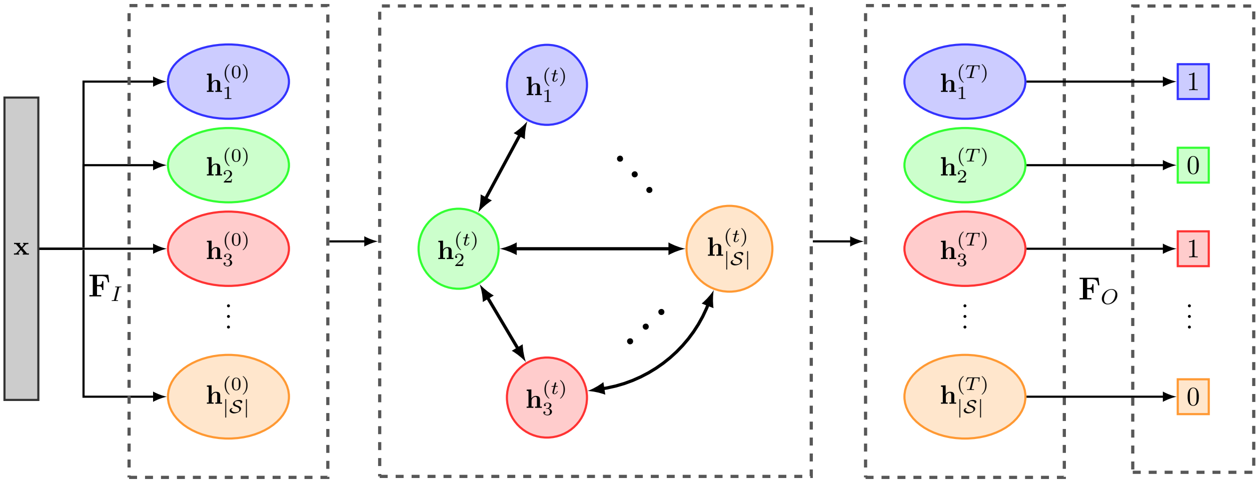

Figure 2. Illustration of structured graph propagation for multi-label classification. Given an input $\mathbf{x}$, we calculate the initial belief state $\mathbf{h}_v^{(0)}$ for each label node. The resulting information is propagated via the observed graph for updating the associated belief states. After propagating $T$ times, the final belief states can be obtained for predicting the final multi-label outputs.

[16] achieves ML-ZSL by exhaustively listing all possible combinations of labels and treating it as a zero-shot classification problem. Recently, [51] considers the separability of relevant and irrelevant tags, proposing a model that learns principal directions for images in the embedding space. Multiple Instance Visual-Semantic Embedding (MIVSE) [39] is another joint embedding method, which uses a region-proposal method to discover meaningful subregions in images and then maps the subregions to their corresponding labels in the semantic embedding space. [17] leverages co-occurrence statistics of seen and unseen labels and learns a graphical model that jointly models the label matrix and the co-occurrence matrix.

# 3. Our Proposed Approach

## 3.1. Notations and Overview

We first define the notations used in this paper. Let $\mathcal{D} = \{(\mathbf{x}^i, \mathbf{y}^i)\}_{i=1}^N$ denote the set of training instances, where $\mathbf{x}^i \in \mathbb{R}^{d_{feat}}$ are $d_{feat}$-dimensional features and $\mathbf{y}^i \in \{0,1\}^{|\mathcal{S}|}$ are the corresponding labels in the label set $\mathcal{S}$. Note that $N$ denotes the number of training instances, while $|\mathcal{S}|$ is the number of seen labels. Given $\mathcal{D}$ and $\mathcal{S}$, the task of multi-label classification is to learn a model such that the label $\hat{\mathbf{y}} \in \{0,1\}^{|\mathcal{S}|}$ of a test instance $\hat{\mathbf{x}} \in \mathbb{R}^{d_{feat}}$ can be predicted accurately.

For ML-ZSL, we have the unseen label set as $\mathcal{U}$, and the goal is to predict the labels in both $\mathcal{S}$ and $\mathcal{U}$ for a test instance $\hat{\mathbf{x}}$. The predicted label is as $\tilde{\mathbf{y}} \in \{0,1\}^{|\mathcal{S}|+|\mathcal{U}|}$, where the first $|\mathcal{S}|$ dimensions are the predictions for the seen label set $\mathcal{S}$, and the bottom $|\mathcal{U}|$ dimensions are for the unseen ones.

Since the images are without the annotation of labels $\mathcal{U}$ during training, ML-ZSL needs to extract the semantic information from the observed label space. In our proposed model, we use distributed word embeddings to represent a class label with a semantic vector. The word embedding is denoted as $\mathbf{W} = \{\mathbf{w}_v\}_{v=1}^{|\mathcal{S}|+|\mathcal{U}|}$, where $\mathbf{w}_v \in \mathbb{R}^{d_{emb}}$ is the word vector representation for label $v$ in $\mathcal{S} \cup \mathcal{U}$, and $d_{emb}$ is the dimension of the word embedding space. In our work, we utilize GloVe [37] as $\mathbf{W}$ with $d_{emb} = 300$.

Our approach is illustrated in Figure 2. We take every label as a node with states in our structured knowledge graph. The initial belief states of these nodes $\mathbf{h}_v^{(0)}$ are first obtained through the input function $\mathbf{F}_I$, and the resulting information is propagated via the structured knowledge graph for updating the belief states. The propagation mechanism from each label node $u$ to a connecting node $v$ is governed by propagation weights $\mathbf{a}_{vu}$, which are produced from the relation function $\mathbf{F}_R^k$. We note that, this relation function takes the label representations $\mathbf{w}_u$ and $\mathbf{w}_v$ as inputs, where $k$ denotes the type of relation between nodes $u$ and $v$ as defined in the knowledge graph. The above propagation and interaction process would terminate after $T$ steps, followed by passing through a output function $\mathbf{F}_O$ to produce the final classification probabilities. In the following subsections, we will give details of how this model is used for ML-ZSL.

## 3.2. Structured Knowledge Graph Propagation in Neural Networks

Inspired by Graph Gated Neural Networks [27, 30], we consider a graph with $|\mathcal{S}|$ nodes, and the propagation model is learned with a gated recurrent update mechanism which is similar to recurrent neural networks. For the task of ML-ZSL, each node $v$ in the graph corresponds to a class label, and there is a belief state vector $\mathbf{h}_v^{(t)} \in \mathbb{R}^{d_{hid}}$ at every time step $t$. Following [30], we set $d_{hid}$ to 5. For ML-ZSL, we cannot simply apply an existing detector as in GSNN to obtain the initial belief states. Instead, we utilize an input function $\mathbf{F}_I(\mathbf{x}, \mathbf{w}_v)$ that takes the input feature $\mathbf{x}$ and the label representation $\mathbf{w}_v$ for each node $v$ as inputs to calculate the initial belief state $\mathbf{h}_v^{(0)}$. The function $\mathbf{F}_I$ is implemented by a neural network.

Next, using the structure of the knowledge graph which encodes the propagation weight matrix $\mathbf{A} \in \mathbb{R}^{|\mathcal{S}|d_{hid} \times |\mathcal{S}|d_{hid}}$, we retrieve the belief states of adjacent nodes and combine the information from adjacent nodes to get an update vector $\mathbf{u}_v^{(t)}$ for each node. The belief states are then updated by a gating mechanism by Gated Recurrent Unit (GRU) with $\mathbf{u}_v^{(t)}$ as the input.

For each class label node $v \in \mathcal{S}$, the propagation recurrence is as follows:

$$\mathbf{h}_v^{(0)} = \mathbf{F}_I(\mathbf{x}, \mathbf{w}_v), \tag{1}$$

$$\mathbf{u}_v^{(t)} = tanh\big(\mathbf{A}_v^\top \left[\mathbf{h}_1^{(t-1)\top} \dots \mathbf{h}_{|\mathcal{S}|}^{(t-1)\top}\right]^\top\big), \tag{2}$$

$$\mathbf{h}_v^{(t)} = GRUCell\big(\mathbf{u}_v^{(t)}, \mathbf{h}_v^{(t-1)}\big), \tag{3}$$

where $\mathbf{A}_v \in \mathbb{R}^{|\mathcal{S}|d_{hid} \times d_{hid}}$ is a submatrix of $\mathbf{A}$ that represents the propagation weight matrix for node $v$ (as detailed in the next subsection). $GRUCell$ is the GRU update mechanism, which is defined as:

$$\mathbf{z}_v^{(t)} = \sigma\big(\mathbf{W}^z \mathbf{u}_v^{(t)} + \mathbf{U}^z \mathbf{h}_v^{(t-1)} + \mathbf{b}^z\big), \tag{4}$$

$$\mathbf{r}_v^{(t)} = \sigma\big(\mathbf{W}^r \mathbf{u}_v^{(t)} + \mathbf{U}^r \mathbf{h}_v^{(t-1)} + \mathbf{b}^r\big), \tag{5}$$

$$\tilde{\mathbf{h}}_v^{(t)} = tanh\big(\mathbf{W}^h \mathbf{u}_v^{(t)} + \mathbf{U}^h (\mathbf{r}_v^{(t-1)} \odot \mathbf{h}_v^{(t-1)}) + \mathbf{b}^h\big), \tag{6}$$

$$\mathbf{h}_v^{(t)} = (1 - \mathbf{z}_v^{(t)}) \odot \mathbf{h}_v^{(t-1)} + \mathbf{z}_v^{(t)} \odot \tilde{\mathbf{h}}_v^{(t)}, \tag{7}$$

where $\mathbf{W}$, $\mathbf{U}$, and $\mathbf{b}$ are learned parameters.

For each time step $t$, the confidence for each label node is obtained by the output function $\mathbf{F}_O$:

$$p_v^{(t)} = \mathbf{F}_O(\mathbf{h}_v^{(t)}), \tag{8}$$

which is implemented by a standard fully-connected neural network. After $T$ time steps for propagation, the final confidences $p_v^{(T)}$ would be obtained.

### 3.3. Learning of the Propagation Matrix

With the gated update mechanism for updating the belief state of each node in a graph, we now address a critical issue that how our model reasons and combines information from adjacent nodes lies in the matrix $\mathbf{A}_v$.

In (2), we see that the update vector $\mathbf{u}_v^{(t)}$ is a weighted combination of the belief states of all other nodes by the propagation parameters in $\mathbf{A}_v$, with each hidden dimension having its own weights. By constraining $\mathbf{A}_v$ to have non-zero weights for the elements that correspond to adjacent nodes and setting weights for non-adjacent nodes to zero, a node would combine information from only relevant nodes that are defined in the structured knowledge graph to obtain the update vector $\mathbf{u}_v^{(t)}$ for updating its own belief state.

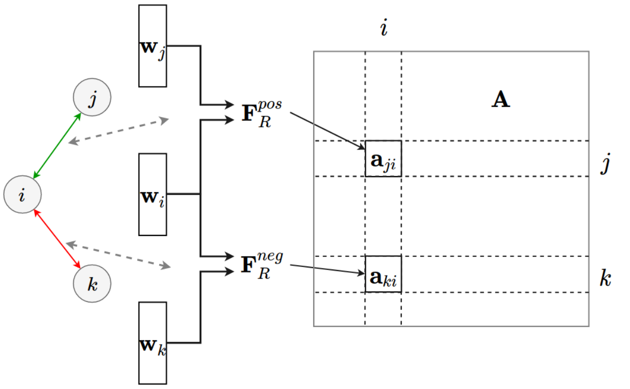

Figure 3. Learning of propagation matrix $\mathbf{A}$ in the semantic space via relation functions $\mathbf{F}_R^k$, with edges defined by the knowledge graph. Note that we only show the propagation from node $i$ outwards, but the matrix $\mathbf{A}$ would be symmetric in practice.

In GSNN, the structured knowledge graph is defined with around 30 relation types. While the elements in $\mathbf{A}$ are learned, the edges of the same relation type are fixed in GSNN. This might limit its practical uses due to only a few relation types can be determined beforehand. In ML-ZSL, it is desirable to exploit finer relation between labels, so the propagation mechanism with the resulting knowledge graph would be sufficiently informative.

To address the above concern, we propose a unique strategy as the propagation weight learning scheme. We combine the informations of word vector into knowledge graphs during propagation stages. Our scheme shares the propagation mechanism for the same relation types while being preferable for ZSL and other practical applications. More precisely, instead of assigning the same propagation weights for edges of the same type/relation, we alternatively assign the same relation function $\mathbf{F}_R^k$ that produces the propagation weights, where $k$ denotes the edge type. Given an edge in edges $\mathcal{E}$ that has edge type $k$, the propagation weights $\mathbf{a}_{vu} \in \mathbb{R}^{d_{hid} \times d_{hid}}$ are determined by:

$$\mathbf{a}_{vu} = \mathbf{F}_R^k(\mathbf{w}_v, \mathbf{w}_u), \tag{9}$$

where $\mathbf{w}_v$ and $\mathbf{w}_u$ are the word vectors for the class label nodes $v$ and $u$. The mechanism for learning propagation weights is illustrated in Figure 3, in which each element of the matrix $\mathbf{a}_{vu}$ is determined by a unique bilinear form from joint embedding of the two associated labels. This allows our model to properly describe relationships between different nodes/relations.

As a final remark, for each edge type $k$, the function $\mathbf{F}_R^k$ learns a mapping from the semantic word embedding

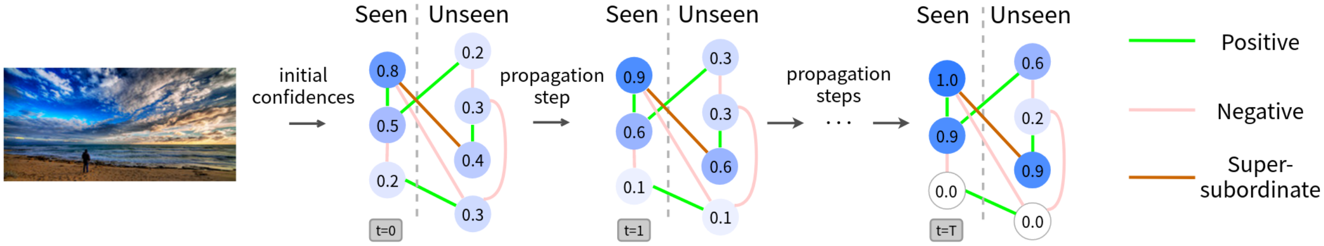

Figure 4. Illustration of information propagation in the knowledge graph. Note that information from a belief state interacts with its adjacent seen/unseen states at each time step based on the relation observed in the knowledge graph. The probabilities represent the outputs of $\mathbf{F}_O$ applied to the belief states at each time step for illustration purposes.

space to the propagation weights, so that the dependency between such relation edges can be modeled accordingly. More importantly, learning from the semantic space allows the aforementioned model to generalize to unseen class labels. Thus, the proposed scheme using relation functions $\mathbf{F}_R^k$ to determine the propagation weights $\mathbf{a}_{vu}$ would be especially preferable for ML-ZSL.

### 3.4. From ML to ML-ZSL

During training, the propagation weight matrix $\mathbf{A}$ can be obtained by forward passing through the relation networks $\mathbf{F}_R^k$, and is then used for information propagation described in (2) to (7). The loss function of our model is a weighted sum of the binary cross-entropy (BCE) of each label node, after the output of network $\mathbf{F}_O$ is observed at each time step. To be more precise, the loss $\mathcal{L}$ is defined as:

$$\mathcal{L} = \frac{1}{N}\frac{1}{|\mathcal{S}|}\sum_{i,v,t}\alpha(t)\Big((y_v^i \log p_v^{(t)} + (1-y_v^i)\log(1-p_v^{(t)}))\Big), \tag{10}$$

where the weights $\alpha(t) = 1/(T-t+1)$ encourage accurate predictions as $t$ increases. During the inference stage of multi-label classification, the final confidences $p_v^{(T)}$ at time step $T$ are used as the predicted outputs.

For ML-ZSL prediction, we extend $\mathbf{A}$ to $\tilde{\mathbf{A}} \in \mathbb{R}^{(|\mathcal{S}|+|\mathcal{U}|)d_{hid}\times(|\mathcal{S}|+|\mathcal{U}|)d_{hid}}$, so that it would encode relations of unseen class labels in the constructed knowledge graph. We also constrain $\tilde{\mathbf{A}}$ so that for edges between $\mathcal{S}$ and $\mathcal{U}$ we only allow propagation from seen to unseen nodes. The update vector $\mathbf{u}_v^{(t)}$ is then calculated from the adjacent nodes for both seen classes $\mathcal{S}$ and unseen classes $\mathcal{U}$. Thus, we have (2) modified as:

$$\mathbf{u}_v^{(t)} = tanh\big(\tilde{\mathbf{A}}_v^\top\,[\mathbf{h}_1^{(t-1)\top}\dots\mathbf{h}_{(|\mathcal{S}|+|\mathcal{U}|)}^{(t-1)}{}^\top]^\top\big), \forall v \in \mathcal{S}\cup\mathcal{U}. \tag{11}$$

The above model is able to calculate the initial belief states for the unseen class labels with $\mathbf{F}_I$, and performs propagation from seen to unseen labels (and also between unseen labels with $\tilde{\mathbf{A}}$ obtained through $\mathbf{F}_R^k$). Finally, the output confidence for each unseen label is derived by $\mathbf{F}_O$. An illustration of the propagation mechanism for ML-ZSL is shown in Figure 4, where the model generalizes from its initial beliefs on seen nodes to the unseen nodes. We note that, during ML-ZSL, our model is also able to produce predictions for the seen class labels in addition to the unseen class labels. Thus, it can be considered for the more challenging task of generalized ML-ZSL.

## 4. Experiments

### 4.1. Building the Knowledge Graph

Before presenting the experimental results, we detail how we built the structured knowledge graph in our model. In our work, we consider WordNet [33] as the source for constructing the knowledge graph, since it is easily accessible and contains rich semantic relationships between different concepts.

We defined 3 types of label relations for the knowledge graph: *super-subordinate*, *positive correlation*, and *negative correlation*. Super-subordinate correlations, also called hyponymy, hypernomy, or ISA relation, is defined and can be directly extracted from WordNet. For positive and negative relations between class labels, label similarities are calculated by WUP similarity [48], followed by thresholding the soft similarities into positive and negative correlations. As for label pairs with similarities between the positive and negative thresholds, or pairs without similarities from WUP similarity, they are viewed as not having any direct relation between them.

In addition, if a pair of labels exhibit super-subordinate relation, we directly apply its resulting dependency in our graph and do not further calculate its positive/negative relation. In the following experiments, we fix the propagation steps on the structured knowledge graph to 5 ($T = 5$).

### 4.2. Datasets and Settings

To evaluate the performance of our model, we consider the following datasets for experiments: NUS-WIDE [9] and Microsoft COCO [28]. For the multi-label classification

| | NUS-81 | | | MS-COCO | | |
|---|---|---|---|---|---|---|
| Method | P | R | F1 | P | R | F1 |
| WSABIE | 30.7 | 52.0 | 38.6 | 59.3 | 61.3 | 60.3 |
| WARP | 31.4 | 53.3 | 39.5 | 60.2 | 62.2 | 61.2 |
| Logistics | 41.9 | 46.2 | 43.9 | 70.8 | 63.3 | 66.9 |
| Fast0Tag | 31.9 | 54.0 | 40.1 | 60.2 | 62.2 | 61.2 |
| Ours | 43.4 | 48.2 | **45.7** | 74.1 | 64.5 | **69.0** |

Table 1. Multi-label classification results on NUS-WIDE with 81 labels and MS-COCO with 80 labels. Results for WSABIE, WARP and Fast0Tag are with $K = 3$.

task we perform experiments on both datasets, while NUS-WIDE is particularly applied for ML-ZSL evaluation.

NUS-WIDE is a web image dataset including 269,648 images and the associated tags from Flickr. For these images, it consists of 1000 noisy labels collected from the web with 81 dedicated ground-truth concepts. We denote these two sets of labels as NUS-1000 and NUS-81, respectively. After collecting all existing images and removing images that do not have any tags, we obtain 90,360 images. We extract 2048-dimensional ResNet-152 [19] feature representations from the images and use them as inputs for the following tasks. We further split the dataset into 75,000 training images, 5,000 validation images and 10,360 test images.

Microsoft COCO (MS-COCO) is a large-scale dataset for object detection, segmentation, and image captioning. We follow the 2014 challenge for data split (i.e., 82783 and 40504 images for training and testing, respectively) with 80 distinct object tags. After removing images without any labels, we split the training set into 78081 training images and 4000 validation images, and the test set is with 40137 images. For all the methods considered in our experiments, we extract and fix 2048-dimensional image features are extracted from ResNet-152.

### 4.3. Multi-Label Classification

We fist consider the conventional multi-label classification tasks for evaluating our proposed model. For comparison, we consider WSABIE [47], WARP [18], and logistic regression (all with the above CNN features) as baseline approaches. We also implement Fast0Tag [51] to compare against models that are designed to handle multi-label classification problems (and later the ML-ZSL tasks).

For testing, since WSABIE, WARP and Fast0Tag predict labels according to the ranking scores of the tags, we choose the top $K$ labels. Following conventional settings, we report results for $K = 3$. As for logistics and our model, every label reports a final confidence for evaluation. Using the validation set, we select a proper probability threshold for predicting labels. Finally, the metrics of precision (P), recall (R) and F1-measure are considered, which are commonly used in previous work.

| | ML-ZSL | | | Generalized | | |
|---|---|---|---|---|---|---|
| Method | P | R | F1 | P | R | F1 |
| Fast0Tag ($K = 3$) | 21.7 | 37.7 | 27.2 | - | - | - |
| Fast0Tag ($K = 10$) | - | - | - | 19.5 | 24.9 | 21.9 |
| Ours w/o Prop. | 31.8 | 25.1 | 28.1 | 24.3 | 23.4 | 23.9 |
| Ours | 29.3 | 31.9 | **30.6** | 22.8 | 25.9 | **24.2** |

Table 2. Results for the ML-ZSL and generalized ML-ZSL tasks on NUS-1000 with 81 unseen labels and 925 seen labels.

Table 1 lists and compares the results for the NUS-81 and MS-COCO datasets. We can see that our model produced comparable performances against baselines. It is worth noting that, since our model is not explicitly designed for solving multi-label but zero-shot learning, the above results sufficiently support the use of our model for multi-label classification. In addition, compared to Fast0Tag, which is designed for ML-ZSL and can also be used in the conventional multi-label setting, our model clearly achieved improved results on both datasets.

We also note that, although Fast0Tag reported higher scores on the recalls on NUS-81, it was not able to produce satisfactory results on the precisions. The discrepancy between precision and recall can also be observed from the results in [51]. Similar remarks can be made for both WSABIE and WARP baselines. A possible explanation is that the number of tags in an image varies across the dataset, and thus simply choosing the top $K$ prediction in terms of ranking scores for every image would not be sufficiently informative. In contrast, logistics and our method applied a more flexible prediction method and were able to achieve more balanced results on precisions and recalls for both datasets.

### 4.4. ML-ZSL and Generalized ML-ZSL

We now report our empirical results on multi-label zero-shot learning (ML-ZSL) using the NUS-WIDE dataset. In order to perform ML-ZSL, we treat labels in NUS-WIDE 81 as the unseen label set $\mathcal{U}$, while the seen label set $\mathcal{S}$ is derived from NUS-1000 with 75 duplicated ones removed and thus results in 925 label classes.

We take Fast0Tag [51] with the same $\mathcal{S}$ and $\mathcal{U}$ as the state-of-the-art ML-ZSL approach for comparisons. We report the results for ML-ZSL with $K = 3$ for Fast0Tag. To further verify the effectiveness of the introduced components in our model, we also conduct controlled experiments in which we have a simplified version without updating the belief vectors via the structured knowledge graph (i.e., Ours w/o Prop.). In other words, for Ours w/o Prop., we set $T = 0$.

Additionally, we consider the challenging task of generalized ML-ZSL task, for which models are trained on seen labels but are required to predict both seen and unseen labels during testing. The experiments are performed on the

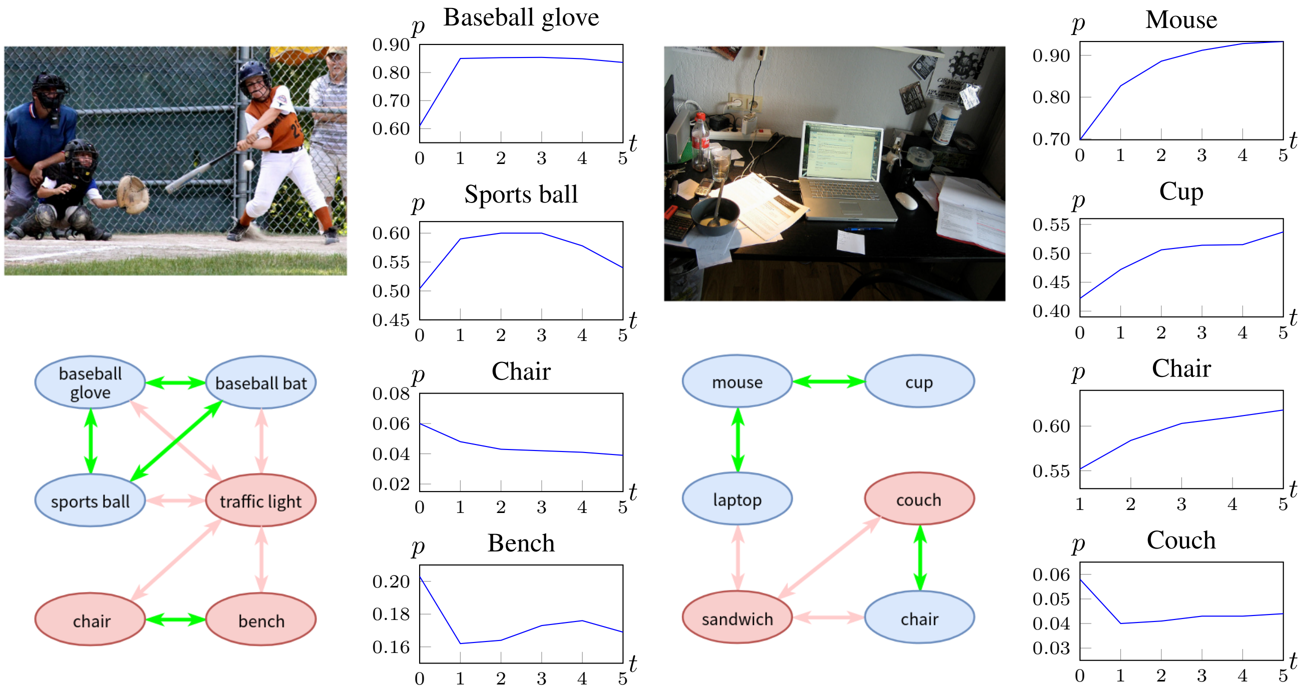

Figure 5. Examples of the constructed knowledge subgraphs and the predicted label probabilities using our proposed method, showing that information propagates across different labels as time step $t$ increases. Note that the blue and red nodes in each subgraph indicate ground truth positive and negative labels, respectively. And, arrows in green or red reflects the corresponding positive or negative relationship.

NUS-WIDE dataset following the ML-ZSL setting, and we report the results of predictions for the $|\mathcal{S}| + |\mathcal{U}| = 1006$ labels. For Fast0Tag under this setting we report $K = 10$, as $K = 3$ will result in low recall due to a large number of tags predicted for each image.

Table 2 lists the results for both the ML-ZSL setting and the generalized ML-ZSL setting. From this table, we see that our model reported satisfactory performances and performed favorably against Fast0Tag. Also, from the ablation tests, we see that the full version of our model was preferable when applying propagation with the knowledge graph. This confirms the effectiveness of this mechanism introduced in our model.

### 4.5. Analysis of Propagation Mechanism

To further evaluate the effectiveness of our method, we visualize the propagation process of our structured knowledge graph in Figure 5, demonstrating how the information transferred in our constructed graph assists in the prediction process. We show the prediction probabilities $p_v^{(t)}$ of several label classes from $t = 0$ to $t = 5$ for the two examples shown in this figure (both are from MS-COCO). These probabilities are obtained from our multi-label classification model. The corresponding knowledge subgraphs are also shown in the figure. From the results, We observe that the first few propagation step affected the prediction probabilities the most, especially for the label nodes that had initial confidence that were closer to the probability threshold. Subsequent propagation steps simply further fine-tuned the probabilities for more accurate predictions.

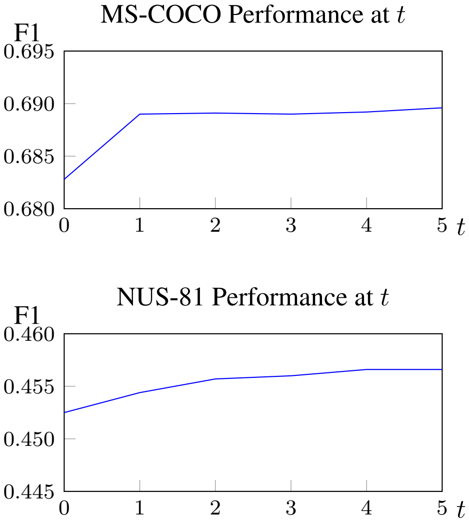

Figure 6. The scores of F1 measure for multi-label classification at different time steps $t$ on MS-COCO and NUS-81.

We also made this similar observation when analyzing the performance of our model at different time steps. We use the probabilities $p_v^{(t)}$ at time step $t$ instead of time step $T$ to obtain predictions and measure the performance on

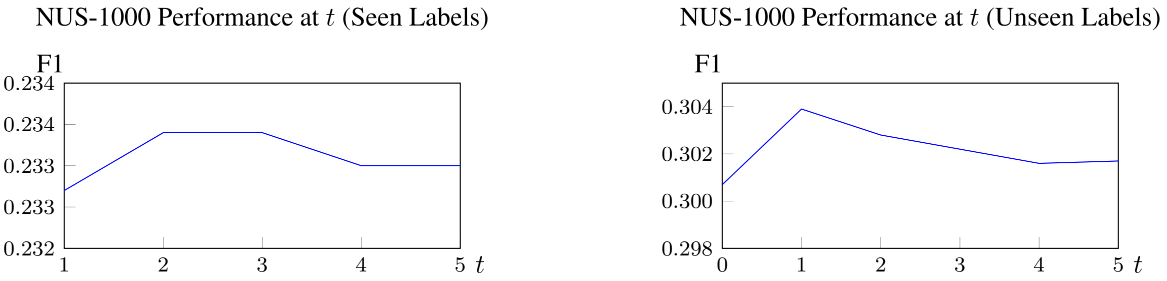

Figure 7. The scores of F1 measure for seen and unseen labels (i.e., generalized ML-ZSL) at different time steps $t$ on NUS-1000.

the testing sets. The results for multi-label classification on MS-COCO and NUS-81 for $t = 1$ to $t = 5$ are shown in Figure 6. In Figures 7, we also observe similar trends for generalized ML-ZSL using NUS-WIDE 1000. In other words, both seen and unseen classes gained from such information propagation across labels, and showed the converged results in a few time steps.

## 5. Conclusion

In this paper, we proposed a unique deep learning framework to approach multi-label learning and multi-label zero-shot learning (ML-ZSL). By incorporating structured knowledge graphs into the learning process, our model leverages different relations defined in the constructed knowledge graph, which allow the exploitation of label dependencies between labels for ML-ZSL. This is similar to how humans utilize learned concept dependencies when recognizing seen and unseen objects of interest. In our experiments, we showed that our proposed model was able to produce satisfactory performance on the standard task of multi-label classification, and performed favorably against baseline and state-of-the-art approaches on the challenging problem of ML-ZSL.

**Acknowledgments** This work was supported in part by the Ministry of Science and Technology of Taiwan under grant MOST 107-2634-F-002-010.

## References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 819–826, 2013.

[2] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2927–2936, 2015.

[3] K. Balasubramanian and G. Lebanon. The landmark selection method for multiple output prediction. In *ICML*. icml.cc / Omnipress, 2012.

[4] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha. Synthesized classifiers for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5327–5336, 2016.

[5] S. Changpinyo, W.-L. Chao, and F. Sha. Predicting visual exemplars of unseen classes for zero-shot learning. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 3496–3505. IEEE, 2017.

[6] Y.-N. Chen and H.-T. Lin. Feature-aware label space dimension reduction for multi-label classification. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1529–1537. Curran Associates, Inc., 2012.

[7] W. Cheng, E. Hllermeier, and K. J. Dembczynski. Bayes optimal multilabel classification via probabilistic classifier chains. In J. Frnkranz and T. Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 279–286. Omnipress, 2010.

[8] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8), 2014*, 2014.

[9] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, page 48. ACM, 2009.

[10] 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 *European Conference on Computer Vision*, pages 48–64. Springer, 2014.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[12] Z. Deng, A. Vahdat, H. Hu, and G. Mori. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4772–4781, 2016.

[13] N. Ding, J. Deng, K. P. Murphy, and H. Neven. Probabilistic label relation graphs with ising models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1161–1169, 2015.

[14] A. Farhadi, I. Endres, D. Hoiem, and D. A. Forsyth. Describing objects by their attributes. In *CVPR*, pages 1778–1785. IEEE Computer Society, 2009.

[15] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. DeViSE: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013.

[16] Y. Fu, Y. Yang, T. M. Hospedales, T. Xiang, and S. Gong. Transductive multi-label zero-shot learning. In *BMVC*, 2014.

[17] A. Gaure, A. Gupta, V. K. Verma, and P. Rai. A probabilistic framework for zero-shot multi-label learning. In *The Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.

[18] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe. Deep convolutional ranking for multilabel image annotation. *CoRR*, abs/1312.4894, 2013.

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[20] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[21] H. Hu, G.-T. Zhou, Z. Deng, Z. Liao, and G. Mori. Learning structured inference neural networks with label relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2960–2968, 2016.

[22] D. Jayaraman and K. Grauman. Zero-shot recognition with unreliable attributes. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3464–3472. Curran Associates, Inc., 2014.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[24] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 951–958. IEEE, 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, Mar. 2014.

[26] J. Lei Ba, K. Swersky, S. Fidler, and R. salakhutdinov. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[27] Y. Li, D. Tarlow, M. Brockschmidt, and R. S. Zemel. Gated graph sequence neural networks. *CoRR*, abs/1511.05493, 2015.

[28] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, Zrich, 2014. Oral.

[29] Z. Lin, G. Ding, M. Hu, and J. Wang. Multi-label classification via feature-aware implicit label space encoding. In *International Conference on Machine Learning*, pages 325–333, 2014.

[30] K. Marino, R. Salakhutdinov, and A. Gupta. The more you know: Using knowledge graphs for image classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[31] T. Mensink, E. Gavves, and C. G. Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[32] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

[33] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, Nov. 1995.

[34] J. Nam, J. Kim, I. Gurevych, and J. Fürnkranz. Large-scale multi-label text classification - revisiting neural networks. *CoRR*, abs/1312.5419, 2013.

[35] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. In *International Conference on Learning Representations (ICLR)*, 2014.

[36] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell. Zero-shot learning with semantic output codes. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1410–1418. 2009.

[37] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[38] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333, Jun 2011.

[39] Z. Ren, H. Jin, Z. Lin, C. Fang, and A. Yuille. Multiple instance visual-semantic embedding. In *Proceeding of the British Machine Vision Conference (BMVC)*, 2017.

[40] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

[41] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[42] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943, 2013.

[43] F. Tai and H.-T. Lin. Multilabel classification with principal label space transformation. *Neural Computation*, 24(9):2508–2542, 2012. PMID: 22594831.

[44] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *Int J Data Warehousing and Mining*, 2007:1–13, 2007.

[45] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu. CNN-RNN: a unified framework for multi-label image classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[46] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan. CNN: single-label to multi-label. *CoRR*, abs/1406.5726, 2014.

[47] J. Weston, S. Bengio, and N. Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI*, volume 11, pages 2764–2770, 2011.

[48] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, 1994.

[49] C.-K. Yeh, W.-C. Wu, W.-J. Ko, and Y.-C. F. Wang. Learning deep latent space for multi-label classification. In *AAAI*, pages 2838–2844, 2017.

[50] M.-L. Zhang and Z.-H. Zhou. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1338–1351, Oct 2006.

[51] Y. Zhang, B. Gong, and M. Shah. Fast zero-shot image tagging. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.