Zero-Shot Learning in the Presence of Hierarchically Coarsened Labels

Colin Samplawski[1], Erik Learned-Miller[1], Heesung Kwon[2], Benjamin M. Marlin[1]
[1]University of Massachusetts Amherst, [2]US Army Research
{csamplawski, elm, marlin}@cs.umass.edu, heesung.kwon.civ@mail.mil

Abstract

Zero-shot image classification leverages side information including label attributes and semantic class hierarchies to transfer knowledge about fine-grained training classes to fine-grained zero-shot classes. In this paper, we consider the problem of zero-shot learning of fine-grained classes given a mixture of images with fine-grained and coarsened labels. We show how probabilistic hierarchical classification models can be used to simultaneously accommodate fine and coarse-grained labels in the zero-shot learning setting. We show that this approach is robust even to significant levels of coarsening.

1. Introduction

Zero-shot learning (ZSL) is one of the most extreme forms of learning from limited labeled data. It enables predicting that test images belong to classes for which no labeled training images are available. ZSL methods accomplish this by introducing a source of side information about classes such as class attribute vectors [7, 3, 17] or a semantic class hierarchy [1, 16, 10]. However, current ZSL methods focus on the problem of learning to transfer knowledge about fine-grained training classes to other fine-grained zero-shot classes using this side information.

In this paper, we consider the problem of zero-shot learning of fine-grained classes where some instances have *hierarchically coarsened* labels. A label is *coarsened* when it corresponds to a set of possible classes instead of a single class [9]. We define a label to be *hierarchically coarsened* if the set of classes it specifies correspond to a sub-tree within a semantic class hierarchy. For example, an instance may be labeled "Dog" instead of an exact breed of dog to indicate that the true label of the instance is only known to lie somewhere within the sub-tree rooted at the "Dog" class.

This work is motivated by the observation that hierarchically coarsened labels are typically easier and cheaper to obtain than exact, fine-grained labels (e.g., it is much easier for a non-expert to recognize that an image contains a dog than it is to specify the precise breed of dog). This makes the integration of large-scale coarsely labeled data sets with smaller-scale fine-grained labeled data sets a potentially attractive zero-shot learning strategy.

We present a principled approach to zero-shot learning that can leverage coarsened labels using a conditional random field (CRF)-based probabilistic hierarchical classifier [12] whose structure is defined using a given semantic class hierarchy. We use a zero-shot parameter prediction approach based on graph convolutional networks (GCNs) [11, 16, 10] to predict the CRF model parameters for zero-shot classes based on parameters for the training classes. This approach significantly generalizes the recent state-of-the-art ZSL method of [16], which is based on parameter prediction applied to a flat multi-class logistic regression model. As a further contribution, we propose two new benchmark data sets for this problem based on the ImageNet data set [14]. Our results show that the proposed approach can robustly incorporate coarsely-labeled training data, outperforming baseline methods even in the presence of high volumes of coarsened labels.

2. Approach

Notation: Let $\mathcal{H} = (\mathcal{Y}, \mathcal{R})$ be the given semantic hierarchy where $\mathcal{Y}$ is the set of all possible classes and $\mathcal{R}$ is a set of "is-a" relations between classes. If $(y, y') \in \mathcal{R}$ for $y, y' \in \mathcal{Y}$, we say that y' is a sub-class of y. The relations in $\mathcal{R}$ are assumed to form a directed tree with a root class y_0 that is a super-class of all other classes (e.g., the "entity" class). We define $\mathcal{Y}_L$ to be the set of all leaf classes of $\mathcal{H}$. These are the most fine-grained classes in $\mathcal{Y}$. We define $\mathcal{Y}_I$ to be the internal classes of $\mathcal{H}$.

In the zero-shot setting, the fine-grained classes in $\mathcal{Y}_L$ can be further partitioned into two sets: $\mathcal{Y}_{zs}$ and $\mathcal{Y}_{tr}$. $\mathcal{Y}_{zs}$ are the zero-shot classes for which by definition we have no labeled training images. $\mathcal{Y}_{tr}$ is the complementary set of fine-grained classes for which we have labeled training images. We assume that every class in $\mathcal{Y}_I$ has at least one descendent in $\mathcal{Y}_{tr}$. We define $\mathcal{A}$ to be a set containing label attribute vectors $\mathbf{a}_y$ for all classes in $y \in \mathcal{Y}$.

Under general label coarsening [9], the labels $z \subset \mathcal{Y}_L$ correspond to a subset of fine-grained classes. In this work,

we consider hierarchical coarsening only. We denote the set of classes in the sub-tree of $\mathcal{H}$ rooted at class y by $S(y)$. The set of all hierarchically coarsened labelings is then defined to be $\mathcal{Z} = \{S(y)|y \in \mathcal{Y}_I \cup \mathcal{Y}_{tr}\}$. Finally, Let $\mathcal{D}_{tr} = \{(\mathbf{x}_n, z_n)|1 \leq n \leq N\}$ be a data set of training instances where $\mathbf{x}_n$ is a training image and $z_n \in \mathcal{Z}$ is the corresponding coarsened class label.

Zero-Shot Learning Problem: Given $\mathcal{H}$, $\mathcal{A}$, and $\mathcal{D}_{tr}$, our goal is to learn a probabilistic classifier $P(Y = y|\mathbf{X} = \mathbf{x})$ that is defined over the set of all fine-grained classes $\mathcal{Y}_L$. If all of the coarsened labels in $\mathcal{D}_{tr}$ are singleton sets such that $z_n = \{y_n\}$ for $y_n \in \mathcal{Y}_{tr}$, then this problem reduces to the standard zero-shot learning scenario where we seek to transfer knowledge about fine-grained training classes to fine-grained zero-shot classes.

However, when the coarsened labels are mix of singleton and non-singleton sets, then we obtain a more general zero-shot learning problem where we seek to transfer knowledge from a mixture of fine and coarsely labeled images to fine-grained zero-shot classes. This more general zero-shot learning problem is the focus of this work. We now present a principled probabilistic model and zero-shot learning framework for addressing this problem.

Probabilistic Hierarchical Classification: The classification model that we use in this work is an instance of the CRF model family [12]. We define a joint probability distribution over all classes in $\mathcal{H}$ conditioned on an input feature vector $\mathbf{x}$. The CRF contains one binary label variable per class in $\mathcal{H}$. We denote a full binary labeling of the hierarchy by the vector-valued random variable $\mathbf{Y}$ and a realization of this variable by $\mathbf{y} \in \{0, 1\}^{|\mathcal{Y}|}$. We introduce the notation $\mathbf{y}(c)$ to refer to the label vector with ones on the unique shortest path from class c to the root and zeros elsewhere. The conditional probability distribution induced by this CRF over the vector of label variables $\mathbf{y}$ given an input $\mathbf{x}$ is specified through the energy function E_θ as shown below.

$$p_\theta(\mathbf{y}|\mathbf{x}) = \frac{\exp(-E_\theta(\mathbf{y}, \mathbf{x}))}{\sum_{c \in \mathcal{H}} \exp(-E_\theta(\mathbf{y}(c), \mathbf{x}))}, \quad (1)$$

$$E_\theta(\mathbf{y}, \mathbf{x}) = \phi^G(\mathbf{y}) + \sum_{c \in \mathcal{H}} \mathbf{y}(c) \mathbf{w}_c^T \phi^F(\mathbf{x}) \quad (2)$$

The model parameters $\theta = \{\mathbf{w}_c|c \in \mathcal{Y}\}$ correspond to the set of weights vectors $\mathbf{w}_c$ associated with each class c. ϕ^F is a feature function mapping input images to feature vectors. The global factor $\phi^G(\mathbf{y})$ ensures that joint label configurations $\mathbf{y}$ respect the "is-a" semantics of the label hierarchy. In particular, the only valid label configurations are those of the form $\mathbf{y}(c)$ for $c \in \mathcal{Y}$. The CRF is explicitly normalized over this set of valid joint labelings, whose cardinality is exactly $|\mathcal{Y}|$.

Given the above definitions, we can easily construct the conditional probability associated with a coarsened label $z \subset \mathcal{Z}$ by marginalizing over the paths from the classes in z to the root of $\mathcal{H}$ as shown in Equation 3. The model is learned by maximizing the conditional likelihood function in Equation 4.

$$p_\theta(z|\mathbf{x}) = \sum_{c \in z} p_\theta(\mathbf{y}(c)|\mathbf{x}) \quad (3)$$

$$\mathcal{L}(\theta; \mathcal{D}_{tr}) = \sum_{n=1}^{N} \log p_\theta(z_n|\mathbf{x}_n) \quad (4)$$

Zero-Shot Parameter Prediction with GCNs: To enable zero-shot prediction using a CRF model we follow a similar approach to [16, 10] which infer classifier parameters for zero-shot classes using a GCN. We first remove the zero-shot classes from $\mathcal{H}$ forming a restricted hierarchy $\mathcal{H}_{-zs}$. We then estimate maximum conditional likelihood parameters $\hat{\mathbf{w}}_c^{MLE}$ for the classes in $\mathcal{H}_{-zs}$. Next, we apply a GCN-based parameter prediction approach to predict the full set of CRF class weights. Specifically, the GCN learning procedure takes as input the class attribute vectors $\mathbf{a}_c$ associated with all classes c in $\mathcal{H}$ and the graph defined by the semantic hierarchy $\mathcal{H}$ and uses them to predict the CRF model parameters $\hat{\mathbf{w}}_c$ associated with all classes.

A GCN consists of layers of the form: $\mathbf{H}_i = g(\mathbf{A}\mathbf{H}_{i-1}^T\mathbf{V}_i)$ where $\mathbf{A}$ is the normalized adjacency matrix of the (undirected) semantic hierarchy $\mathcal{H}$, $\mathbf{V}_i$ is the set of weights for layer i, and g is an activation function. $\mathbf{H}_0$ is the input matrix of class attribute vectors $\mathbf{a}_c$ for each class c. After a series of such layers, the model predicts output vectors $\hat{\mathbf{w}}_c$ for all classes c in $\mathcal{H}$. To learn the GCN, we minimize the squared loss function $\sum_{c \in \mathcal{Y}_{tr}} ||\hat{\mathbf{w}}_c^{MLE} - \hat{\mathbf{w}}_c||_2^2$.

After training the GCN, we use it to predict parameters $\hat{\mathbf{w}}_c^{GCN}$ for every class in $\mathcal{H}$. While [16, 10] replace the MLE parameters $\hat{\mathbf{w}}_c^{MLE}$ with the GCN-inferred parameters $\hat{\mathbf{w}}_c^{GCN}$ for all classes, we instead retain the MLE parameters $\hat{\mathbf{w}}_c^{MLE}$ for training class and internal nodes and use the $\hat{\mathbf{w}}_c^{GCN}$ GCN-inferred parameters for zero-shot classes only, which we find improves performance.

As discussed by [16], the GCN model is easier to train when the outputs of the GCN are L2-normalized. In general, the MLE parameters of the base model do not have unit norm, so the GCN-predicted parameters must be scaled in order to achieve reasonable performance. To that end, we calculate the final predicted parameters using the formula: $\hat{\mathbf{w}}_c^* = [c \in \mathcal{Y}_{zs}] \cdot \gamma \cdot \hat{\mathbf{w}}_c^{GCN} + [c \notin \mathcal{Y}_{zs}] \cdot \hat{\mathbf{w}}_c^{MLE}$. We set the scale parameter γ to the mean of the 2-norm of the MLE parameter vectors $\hat{\mathbf{w}}_c^{MLE}$ for $c \in \mathcal{Y}_{tr}$.

Implementation Details: In the experiments that follow, we use a pre-trained ResNet101 [8] model trained to classify the 1,000 ImageNet classes as the image feature function $\phi^F(\mathbf{x})$, resulting in a $D = 2048$ dimensional feature vector. For the label attribute vectors $\mathbf{a}_c$, we trained a 300-dimensional Subword Information Skip Gram model [2] on all of English Wikipedia using the fastText library. Hyper-

Dataset	Training Classes	Training Images	Zero-Shot Classes	Zero-Shot Images	Superclasses
Small Hierarchy	53	∼69,000	41	∼47,000	8
Large Hierarchy	1,000	∼1,300,000	897	∼1,100,000	376

Table 1. Statistics for benchmark hierarchies. Training and zero-shot classes are leaves. Superclasses are non-leaf classes. Labeled images are available for training and zero-shot classes only. Labeled zero-shot class images are used for testing only.

parameters were chosen using a grid-search based selection process which used a simulated zero-shot partitioning of the training classes. In all cases we used ReLU activation functions and dropout regularization. Optimization of the GCN models was performed using Adam and the CRF models were trained using SGD with momentum.

3. Benchmark Datasets

In this section, we present the hierarchical zero-shot benchmark data sets used in our evaluation. We construct benchmarks using the ImageNet dataset [14].

Existing Benchmarks: The most commonly used ImageNet subset is the 1,000 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) classes. However, the full ImageNet graph is made up of about 32,000 classes of which 21,000 have images. One common approach to generate zero-shot test sets from these 21,000 classes is use the distance from the 1,000 ILSVRC training classes. For example, popular sets are the "2-hops" and "3-hops" test sets, which consist of sets of classes which are 2 and 3 tree hops away from the 1,000 training classes, respectively [7].

Unfortunately, little attention has been paid to the structure of these commonly used test sets. For example, the ILSVRC 2-hops set consists almost entirely of children or parents of the ILSVRC training nodes. Furthermore, the ILSVRC 3-hops set often exhibits a nested structure where zero-shot classes are parents of other zero-shot classes. This is especially problematic for the approaches of [16, 10] since their methods output a flat classifier over the seen and unseen classes. They are therefore implicitly learning a flat classifier over a set of nested classes. We instead propose to restrict the zero-shot classes to be leaf classes in a semantic hierarchy $\mathcal{H}$ extracted from WordNet.

Small Benchmark Hierarchy: As a simplified baseline, we extend the 65 node hierarchy presented by [5]. We change this hierarchy by removing the four leaf nodes that do not appear in the 1,000 ILSVRC classes. We do this so that we can use a pre-trained ResNet model without the need to fine-tune. Furthermore, we choose 41 classes which are siblings of the training leaves in the larger WordNet hierarchy to use as zero-shot classes.

Large Benchmark Hierarchy: We also construct a large benchmark semantic hierarchy for ImageNet that includes all 1,000 of the ILSVRC classes. We start with the full WordNet graph of nouns [13], of which ImageNet is a subgraph. We then iteratively remove leaf nodes from the graph until the only remaining leaves are the 1,000 ILSVRC classes. This results in a graph which contains non-tree edges (i.e., some nodes have more than one parent). We enforce a tree structure by running Chu-Liu/Edmonds' algorithm [6], which finds an optimal spanning tree given a general graph as input. We then contract all nodes which have exactly one child. We generate a set of 897 zero-shot classes that (1) have at least 1000 images and (2) are sibling nodes of the 1,000 ILSVRC classes. Table 1 displays summary statistics for these two benchmark hierarchies.

4. Related Work

In terms of hierarchical classification models, the hierarchy and exclusion (HEX) graph model framework of [4] is the most closely related work. The CRF model that we leverage is equivalent to that of [4] when only hierarchy edges are included. This results in efficient linear-time marginalization and normalization. [4] consider learning HEX graph models from coarsened labels, but do not consider the zero-shot learning problem as defined in this work.

In terms of zero-shot learning, the most closely related work is that of [16], which can be viewed as a simplified special case of our approach. [16] use a GCN to predict the parameters of a flat, multi-class logistic regression classifier over fine-grained classes. The hierarchical CRF model that we use is the key to enabling principled incorporation of hierarchically coarsened labels, which [16] do not consider. This strictly generalizes the use of a multi-class logistic regression classifier. As noted earlier, we also retain the MLE parameters for the training classes and explicitly re-weight the GCN parameter predictions to improve performance.

The recent approach of [10] builds upon [16] by designing a new GCN architecture for this parameter prediction problem. At a high level the approach is the same, with only the GCN changing. In our initial experiments, we found that the approaches of [10] and [16] performed similarly.

Finally, generative models including VAEs have seen increasing use in zero-shot learning. The cross-alignment and distribution-alignment variational autoencoder (CADA-VAE) [15] learns a separate VAE for each modality of the data (e.g., images and attributes) and uses a cross-alignment loss and a Wasserstein distance to align the latent spaces of the VAEs. A classifier is then trained in the aligned latent space using encodings of images and attributes. CADA-VAE is a strong a baseline for fine-grained zero-shot classification, but does not leverage hierarchical information.

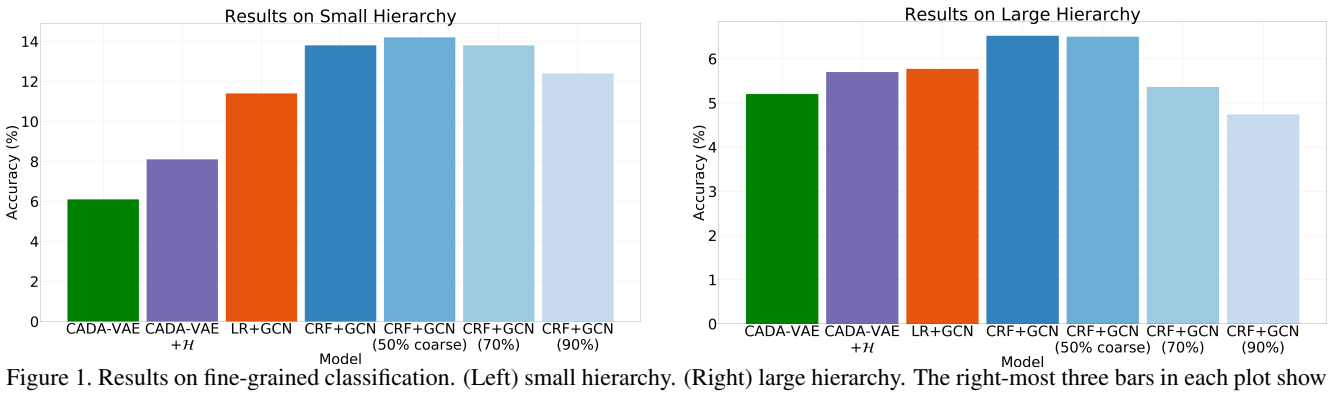

Figure 1. Results on fine-grained classification. (Left) small hierarchy. (Right) large hierarchy. The right-most three bars in each plot show the CRF+GCN under increasing levels of coarsening. The left-most four bars show results without coarsening.

5. Experiments and Results

In this section, we present experiments and results.

Experimental Details: We use logistic regression combined with graph convolutional networks (LR+GCN, which uses our modified parameter prediction procedure) and CADA-VAE as a baselines. We also use a modification of CADA-VAE that includes an additional semantic hierarchy-based embedding of each class (CADA-VAE+$\mathcal{H}$). The hierarchy-based class embeddings are created from the class hierarchy following the approach of similar to [1] which constructs a node embedding as a normalized vector of node adjacency information. We compare to the proposed combination of conditional random field models with graph convolutional networks (CRF+GCN). All models used the same image features and label attribute embeddings.

We perform experiments with both the small and large benchmark hierarchies. All test results use the more challenging generalized zero-shot prediction procedure where the true test labels are in $\mathcal{Y}_{zs}$, but the model makes free predictions over the full set of leaf classes $\mathcal{Y}_L = \mathcal{Y}_{zs} \cup \mathcal{Y}_{tr}$ [3].

For the larger benchmark hierarchy, the best performing architecture for both the CRF and logistic regression was a six layer GCN with output dimensions: $2048 \rightarrow 2048 \rightarrow 1024 \rightarrow 1024 \rightarrow 512 \rightarrow D$. For the smaller hierarchy, the best performing GCN for the CRF was 3 layers with output dimensions $2048 \rightarrow 1024 \rightarrow D$. Lastly, the best performing GCN for the logistic regression model was a 2-layer GCN with output dimensions $2048 \rightarrow D$.

Experiment 1: Fine-Grained Prediction from Fine-Grained Labeled Training Instances: We first consider the case of fine-grained prediction within the leaves of the hierarchy $\mathcal{H}$ when learning using only fine-grained training labels. The results are shown in Figure 1 for the small and large benchmarks (left four bars of each figure). We see that the LR+GCN yields significant improvement relative to CADA-VAE without hierarchical information, which is consistent with the ImageNet results of [15]. Further, we see that adding hierarchical class embeddings to CADA-VAE leads to improved performance on both hierarchies, confirming the result of [1]. The CRF+GCN method, which makes more extensive use of the hierarchy through the CRF model, makes further improvements over LR+GCN and CADA-VAE on both problems.

Experiment 2: Fine-Grained Prediction from Coarsely Labeled Training Instances: In this experiment, we simulate a dataset in which varying amounts of data have coarsened labels. As in [4], we consider re-labeling varying fractions of the training instances with the parent label of their true class during training. We note that only the CRF+GCN method can make use of the coarsened labels during training (right 3 bars of each figure). We learn it under coarsening, assess its fine-grained generalized zero-shot test accuracy, and compare to the baseline methods *learned on all training data without coarsening*. Remarkably, the CRF+GCN is able to maintain stable performance with up to 50% of training labels coarsened. Further, we can see that the CRF+GCN method with up to 90% coarsening out-performs LR+GCN with no coarsening on the small benchmark. Similarly, the CRF+GCN method with up to 50% coarsening maintains its advantage over the LR+GCN method with no coarsening on the large benchmark.

6. Conclusions

We have presented a novel approach to zero-shot learning that is able to effectively leverage large volumes of coarsely labeled training instances by combining the strength of a hierarchical probabilistic classifier with a graph convolutional parameter prediction approach that also leverages class hierarchy information. Our results show that our proposed model out-performs prior state-of-the-art zero-shot image classification methods in the standard setting, while maintaining very strong zero-shot performance in the presence of significant volumes of coarsened labels.

Acknowledgements

This work was partially supported by the US Army Research Laboratory under cooperative agreement W911NF-17-2-0196. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the US government.

Author Erik Learned-Miller was sponsored by the AFRL and DARPA under agreement number FA8750-18-2-0126. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the AFRL and DARPA or the U.S. Government.

References

[1] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, pages 2927–2936, 2015.

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. In *TACL*, volume 5, pages 135–146, 2017.

[3] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *ECCV*, pages 52–68, 2016.

[4] Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In *ECCV*, 2014.

[5] Jia Deng, Jonathan Krause, Alexander C Berg, and Li Fei-Fei. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In *CVPR*, 2012.

[6] Jack Edmonds. Optimum Branchings. *Journal of Research of the National Bureau of Standards*, 71B(4):233, 1967.

[7] Andrea Frome, Greg S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129, 2013.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[9] Daniel F Heitjan and Donald B Rubin. Ignorability and coarse data. *The annals of statistics*, pages 2244–2253, 1991.

[10] Michael Kampffmeyer, Yinbo Chen, Xiaodan Liang, Hao Wang, Yujia Zhang, and Eric P Xing. Rethinking knowledge graph propagation for zero-shot learning. In *CVPR*, pages 11487–11496, 2019.

[11] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2016.

[12] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289, 2001.

[13] George A Miller. Wordnet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[14] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet: large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[15] Edgar Schönfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *CVPR*, pages 8247–8255, 2019.

[16] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *CVPR*, pages 6857–6866, 2018.

[17] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning - the good, the bad and the ugly. In *CVPR*, pages 3077–3086, 2017.