INTEGRATING KNOWLEDGE GRAPH INFORMATION FOR FEW-SHOT LEARNING WITH GRAPH NEURAL NETWORKS

Ethan Shen
Department of Computer Science
Stanford University
ezshen@stanford.edu

ABSTRACT

Traditional machine learning still falls short of humans at tasks in data constrained regimes as seen in zero and few shot learning. Can we mimick humans' incredible ability to generalize new concepts with little to no supervision by integrating contextual graph-based information? Here, we extend a formulation of graph neural networks for the few-shot classification task by introducing a new loss function regularizing the learned graph Laplacian to approach ground truth knowledge graph distances of the *class* labels. We demonstrate the efficacy of our model by integrating WordNet graph information to the few-shot Mini-Imagenet classification task. Ablation studies suggest our model gives an accuracy boost to state of the art graph neural network models for this task when used as an additional module.

1 INTRODUCTION

In the traditional supervised machine learning framework for classification, the output classes are considered as distinct semantic categories, often represented as one-hot encoding vectors, with no consideration of inter-category information. In the context of few-shot or zero-shot machine learning, this framework for supervised training fails miserably Norouzi et al. (2013), as models trained on a class-limited training set cannot generalize effectively to unseen or barely-seen classes. Though there have been continued efforts in collecting larger data corpora with broader coverage of concepts and categories, the few and zero-shot learning problem still remains a significant hurdle, and improvements would significantly increase real-world applicability of models especially for data-starved regimes. Recent advances seek to solve this problem by taking advantage of a simple fact: our world, and how we categorize concepts, has structure. From a human intuition perspective, how can we tell that we are looking at a zebra, even though we have never seen one before? Perhaps one way is by reading somewhere that a zebra is a horse-like animal with black and white stripes. Efforts in this field seek to capture this natural intuition.

One approach is to enforce explicit structural relations from existing ground truth knowledge, which can be represented by graphs. Here, we summarize our work as follows:

1. We review the work of Garcia & Bruna (2017) and cast few-shot learning as a supervised message-passing task which is trained as a Graph Neural Network (GNN).
2. We introduce a new graph distance loss function which integrates information from the ground truth graph relationships between classes.
3. We demonstrate improved performance on the Mini-Imagenet use case, and run ablation studies and error analysis to show that our formulation is most effective when used in combination with the model proposed by Garcia & Bruna (2017).

In the following sections, we first review related works in Section 2. We then provide a mathematical formulation of the few-shot problem in Section 3. Section 4 describes the model. Finally, Sections 5 and 6 describe the Mini-Imagenet dataset, and present experimental results and evaluations. Though

the backbone of our study is based on that of Garcia & Bruna (2017), all work in Sections 4.4, 4.5, 5, and 6 is novel.

2 PRIOR WORK

Given the difficulty of the zero and few-shot learning task, a significant line of work has been developed in integrating outside contextual information.

For example, Socher et al. (2013) captures intuitive "contextual knowledge" by mapping images into the semantic space of word labels that is learned by a neural network model, with word vectors learned from a large, unsupervised text corpus. Another approach is to take convex combinations of the output semantic embeddings from the training set classes to produce embeddings for unseen ones Norouzi et al. (2013).

Here, we review prior work on two strands of research: improvements in the field on zero and few-shot learning, and graph neural networks applied to this task.

2.1 PROTOTYPICAL NETWORKS FOR FEW-SHOT LEARNING

As model efficacy for image recognition and semantic word embeddings have improved, researchers have built upon the GloVE Pennington et al. (2014) embedding work to also improve zero and few-shot learning models. For example, Snell et al. (2017) introduce a model that embeds images using a deeper CNN model into category space to classify unseen categories, which they dubbed ProtoNets. They hypothesize that there exists an embedding in which points cluster around a single prototype representation for each class. During training, they learn a non-linear mapping from the input space to into an embedding space, and take these category "prototype" to be the mean of its support set in the embedding space. During classification time, the output category is simply the closest anchor point to the input embedding, as defined by some distance metric. In this paper, the authors contribute the following strong and interesting findings:

- During inference time, using squared Euclidean distance instead of the more commonly used Cosine distance greatly improves prediction results. They hypothesize this is because euclidean distance is a Bregman divergence as opposed to Cosine distance. It has been shown for Bregman divergences that the cluster representative achieving minimal distance to its assigned points is the cluster mean formulated by Banerjee et al. (2005), which is exactly the "prototype" point.
- Training schedule matters for zero-shot generalization. Experimentally, the authors note that increasing the number of unseen classes during each episode of training increases accuracy during test time, but maintaining the same number of examples per class during training and testing is generally better.
- Their framework extends naturally to zero shot learning, where the model can learn some embedding of the class metadata (e.g. word vectors) and perform a similar inference procedure.

This concept (including papers which make incremental improvements off this work, such as replacing the distance measure during inference with a learned neural network Sung et al. (2018), is currently the state-of-the-art technique for few-shot learning. Here, some of the most impressive strides come from improvements in deep image embedding architectures. Overall, this work shows that one effective way to solve the few and zero-shot learning problem is to map example embeddings to the semantic embedding space of labels, so we can capture some label information, instead of just using discrete representations, such as one-hot embeddings. One way to capture this label information is to assume some underlying graph structure, which we discuss below.

2.2 ZERO-SHOT RECOGNITION VIA SEMANTIC EMBEDDINGS AND KNOWLEDGE GRAPHS

Recent work by Kipf & Welling (2016) develops an architecture to perform deep learning over graphs. This paper applies the GCN architecture to the zero-shot learning problem. They use a 6-layer Graph Convolutional Network (GCN) model to transfer information (message-passing) between different categories that takes word vector inputs and outputs classifier vectors for different

categories. During inference, these classifier vectors are multiplied with the image embeddings to produce classification scores. Here, the authors constructed a knowledge graph based on NELL in Carlson et al. (2010) and images from NEIL in Chen et al. (2013) for corresponding categories, and show a significant increase in zero-shot learning accuracy compared to baselines. However, the authors did not show experimental results on standard zero-shot learning datasets, such as mini-Imagenet by Vinyals et al. (2016), which makes it difficult to compare against state-of-the-art models, such as by ProtoNet Snell et al. (2017) or RelationNet by Sung et al. (2018). These baseline controls are important to distinguish whether the improvements in performance are due to deep learning model architecture improvement (AlexNet by Krizhevsky et al. (2012) to ResNet by He et al. (2016)) or the GCN.

Overall, the body of work in few-shot learning with graph neural networks is still recent, and is a promising field of exploration. We detail our contributions below.

3 Problem Set-Up

3.1 Few-Shot Learning

Here, we introduce the general problem setup and notations for the few-shot learning problem, following the notation of Garcia & Bruna (2017). In this problem, we are given a small subset of size N of all the classes. For each class, we choose a limited set of K labeled training examples. Given a new example from one of the N classes, can we accurately predict its class? This problem is described as k-shot, N-way learning. Formally, we consider a sets T_i of input output pairs drawn iid from a distribution L of labeled image collections.

$$T = \left\{ \{(x_1, l_1), ..., (x_s, l_s)\}, \{\bar{x}\}; l_i \in \{1, N\}, x_i, \bar{x} \in P_l(\mathbb{R}^D) \right\}$$

where s is the number of labeled training samples and N is the number of classes. $P_l(\mathbb{R}^D)$ denotes a class-specific image distribution over $\mathbb{R}^D$, from which we sample iid training and test examples $x_i, \bar{x}$. When $s = kN$, this formulation corresponds to the k-shot, N-way learning problem we described above. The goal is to predict the $\bar{y} \in \{1, N\}$ associated with $\bar{x}$. Given a meta-training set of $(T_i, \bar{y}_i)_{i \leq L}$ over L splits, we consider the standard supervised learning objective

$$\min_{\Theta} \frac{1}{L} \sum_{i \leq L} \ell(\Phi(T_i; \Theta), \bar{y}_i) \tag{1}$$

for a loss function ℓ, parameters Θ. We will use the model $\Phi(T; \Theta) = p(\bar{y}|T)$ which we formalize below.

4 Model

Here, we describe the main architecture we will use, and explain some intuitions as applied to our use-case.

4.1 Embedding Model

The embedding architecture used for Mini-Imagenet images is formed by a 4 convolutional layers followed by a fully-connected layer resulting in a 128 dimensional embedding. This light architecture is useful for fast prototyping, and its modularity allows replacement with any number of more wider or deeper, including ResNet-50, VGG-16, etc. We use the same embedding model as Garcia & Bruna (2017) to maintain comparability, the architecture is as follows:

$1 \times \{3 \times 3$-conv. layer (64 filters), batch normalization, max pool(2, 2), leaky relu$\}$,
$1 \times \{3 \times 3$-conv. layer (96 filters), batch normalization, max pool(2, 2), leaky relu$\}$,
$1 \times \{3 \times 3$-conv. layer (128 filters), batch normalization, max pool(2, 2), leaky relu, dropout(0.5)$\}$,

$1 \times \{3\times3$-conv. layer (256 filters), batch normalization, max pool(2, 2), leaky relu, dropout(0.5)$\}$,
$1 \times \{$fc-layer (128 filters), batch normalization$\}$.

4.2 Similarity Learning with Graph Propagation

Given a set T of limited training examples, we seek to associate the test example $\bar{x}$ with a label $\bar{y}$ from one of the N classes. One of the key challenges of the few-shot problem is difficulty of learning a metric function to compare input examples in embedding space, and much work has been performed in this space.

We formulate this as a graph propagation problem. The goal is to propagate label information from labeled samples towards the unlabeled query image. For each learning set T, we construct a fully connected undirected graph $G_T = (V, E)$ where nodes $v_i \in V$ correspond to both labeled and unlabeled images. The setup does not specify a fixed $e_{i,j} \in E$. Instead, we seek to learn a similarity measure between each node of the graph, and formulate the few-shot learning problem as a node classification problem where the underlying adjacency matrix is learned. This framework is closely related to the set representation from Vinyals et al. (2016), and we will show that our framework is a generalization of this and other canonical works below.

We initialize the nodes by concatenating the output from our embedding model in Section 4.1, $\psi(x_i) = z_i$ with a one-hot representation of the ground truth example label $h_i = h(l_i)$, to get $v_i = \{z_i, h_i\}$. This gives us node embeddings in $\mathbb{R}^{D+N}$. We use these node embeddings as input to our GCN model below.

4.3 Graph Convolutional Networks

Here, we use Graph Convolutional Networks (GCNs) as formulated by Kipf & Welling (2016) to learn a similarity metric between few-shot examples. GCNs seek to aggregate structural information from graphs, and applied it to perform entity classification. This work builds upon approaches from two general categories: explicit graph Laplacian regularization (Weston et al. (2012)) and graph embedding-based approaches inspired by random walk models, including DeepWalk (Perozzi et al. (2014)) and node2vec (Grover & Leskovec (2016)).

A convolutional propagation rule that operates directly on graphs and generates node embeddings based on local network neighborhoods as shown below:

$$f(H^{(l)}, \hat{A}) = \sigma(\hat{D}^{(-\frac{1}{2})}\hat{A}\hat{D}^{(-\frac{1}{2})}H^{(l)}W^{(l)})$$

where f is a function of H, the previous input hidden layer, and $\hat{A}$ which is a learned adjacency matrix. $\hat{D}$ is a normalized diagonal matrix, and $\hat{A}$ is the normalized adjacency matrix. In our case $H^{(1)} = \{v_1, ...v_s\}$. We seek to learn the weight matrix for each layer $W^{(l)}$. Theoretically, it is motivated from a first-order approximation of spectral graph convolutions. From these hidden layers we can extract a node output $\hat{h}$ which we can feed into our loss function detailed in Section 4.5.

We also seek to learn an Adjacency matrix $\hat{A}$ from node hidden representations. We have a simple formulation:

$$\hat{A}_{i,j} = \text{MLP}(|z_i - z_j|)$$

this adjacency matrix is then fed into the GCN, and backpropagation gradients can be passed through.

4.4 Normalized Graph Distance Laplacian

We seek to learn an appropriate normalized adjacency matrix $\hat{A}$ to maximize the few-shot learning objective in (1). We hypothesize that integrating information from the relationships between Mini-Imagenet classes would improve few-shot generalization. From WordNet, we calculate graph distance functions between each of the N classes, to create a distance matrix $D \in \mathbb{R}^{N\times N}$. More detail can be found in Section 5.2.

We first calculate a diagonal matrix of transmissions of the vertices of the graph $T = \text{diag}(\sum_i D_i)$. We then calculate the normalized graph distance Laplacian L in the following manner:

$$L = I - T^{-1/2} D T^{-1/2}$$

where I is the identity matrix. This is analogous to the usual normalized Laplacian matrix, $L = I - D^{-1/2} A D^{-1/2}$ where A is the weighted adjacency matrix and D is the diagonal degree matrix.

4.5 TRAINING

To train the above model, we can formulate a multi-objective loss function. The first loss is a simple cross-entropy loss across the N classes, evaluated for the GCN output $\bar{h}$.

$$\mathcal{L}_{class} = -\sum_n y_n \log P(\bar{h} = y_n | T)$$

The second loss seeks introduce supervision for the learned adjacency matrix $\hat{A}$ by minimizing the distance between the current graph laplacian $\hat{L} = I - \hat{D}^{-1/2} \hat{A} \hat{D}^{-1/2}$ and the ground truth normalized graph Laplacian calculated from the previous section:

$$\mathcal{L}_{graph} = \left\| \hat{L} - L \right\|_2$$

To train the multi-objective loss function we minimizes the convex combination of the losses according to a hyperparameter α,

$$\mathcal{L} = \alpha \mathcal{L}_{class} + (1 - \alpha) \mathcal{L}_{graph}$$

5 DATASETS

We work with the following datasets for the few-shot task.

5.1 MINI-IMAGENET

This Mini-Imagenet dataset was proposed by Vinyals et al. (2016) derived from the original ILSVRC-12 dataset Krizhevsky et al. (2012). Its complexity is high due to the use of Imagenet images but requires fewer resources than running the entire Imagenet dataset, which makes it more suitable for fast prototyping. In total, there are 100 classes with 600 samples of 84×84 color images per class. These 100 classes are divided into 64, 16, and 20 classes respectively for sampling tasks for meta-training, meta-validation, and meta-test. Each class has around 600 samples.

5.2 WORDNET KNOWLEDGE GRAPH

Inspired by Wang et al. (2018), we sought to integrate the internal WordNet (Fellbaum (1998)) knowledge graph associated with the labels Mini-Imagenet, which includes around 74k nodes and 75k edges. The distribution of graph distances for the classes are shown in Figure 1.

For each of the 100 Mini-Imagenet classes, we calculated its shortest graph distance $d(l_i, l_j)$ to each of the other 100 classes. This graph distance matrix was used for the graph adjacency loss computation in Section 4.4.

6 EXPERIMENTS

6.1 BASELINE MODEL

Here, we describe a simple baseline model for comparison. We take the embedding model described in Section 4.1 and learn a linear layer from the embeddings $z_i \in \mathbb{R}^D$ to an output in $\mathbb{R}^N$, the number of classes. We then take the softmax of this output, and run cross entropy loss.

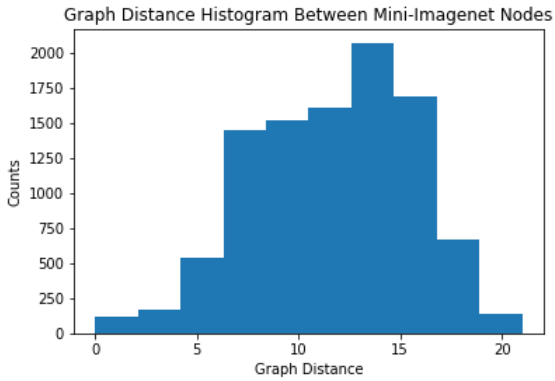

Figure 1: Graph distance on Wordnet graph between each Mini-Imagenet node. Calculated using Stanford SNAP software, Leskovec & Sosič (2016).

6.2 Results and Ablation Study

We ran the experiments with the following settings: $\alpha = 0.1$, batch size is 400, embedding dimension is 128, learning rate is 0.001, and momentum is 0.5. We performed hyperparameter tuning only on the combination parameter α, all other hyperparameters were taken from Garcia & Bruna (2017). We also included ablation study results in the table. Class loss refers to the model trained on only $\mathcal{L}_{class}$ and graph loss refers to the model trained on only $\mathcal{L}_{graph}$.

	1-Shot 5-Way	1-Shot 15-Way
Baseline	19.71%	7.60%
Matching Networks Vinyals et al. (2016)	43.60%*	—
ProtoNet Snell et al. (2017)	46.61%*	—
MAML Finn et al. (2017)	48.70%*	—
GNN Garcia & Bruna (2017)	50.33%*	—
Our GNN (Class Loss)	48.59%	24.00%
Our GNN (Graph Loss)	21.40%	8.40%
Our GNN (Class Loss + Graph Loss)	49.86%	24.18%

Figure 2: Table 1: Few-Shot Learning ablation study on Mini-Imagenet. Starred results are taken from references.

.

Here, we see that marginal gains are made by using incorporating the graph loss term $\mathcal{L}_{graph}$. Ablation analysis shows that the class loss is necessary for any learning to happen, if we only train with graph loss the accuracy returns to baseline levels. However, using a combination of class loss and graph loss we get a large boost in the 1-Shot 5-Way experiment, and smaller boost in the 1-Shot 15-Way experiment. The expected accuracy gain may be a little lower than expected, considering we are including an entire mode of relevant supervision. Our hypothesis is that the bottleneck here actually is that the embedding model is overfitting on the training data, especially due to the small size of labeled data, and the small size of the embedding model itself. Therefore, any focus on learning similarities between samples may not be sufficient because of the low quality of the image embeddings themselves. Indeed, the training accuracy is extremely high (around 70%, data not shown) compared to the validation accuracy (49%). Furthermore, recent work shows that even the baseline model can outperform many of the given comparison models, such as ProtoNet and MAML, given a sufficiently deep embedding model such as ResNet50 Dhillon et al. (2019).

6.3 QUALITATIVE ERROR ANALYSIS

Here, we take a look at the successes and errors that the model made. Overall, the model was especially good at predicting classes where the object images had similar viewpoints. For example, it was especially good at predicting combination locks (Figure 3, middle), even after only seeing one of them during training. This may be because most pictures of combination locks prominently feature the locks themselves, so it was easier to generalize across examples in the class.

The model was unable to discern very specific classes, such as the difference between the ”dung beetle” and the ”rhinocerous beetle” (Figure 4, left). Both classes are black beetles, often shiny in pictures. These sorts of errors may be mitigated by a stronger embedding network, as discussed in the section above.

Figure 3: Randomly sampled successful examples. The predicted classes were ”catamaran”, ”combination lock”, and ”garbage truck” from left to right.

.

Figure 4: Randomly sampled failure examples. The predicted classes were ”dung beetle” and ”ski” from left to right.

7 CONCLUSION

Few/zero-shot learning efficacy is crucial for applying machine learning in the real world, where data is unlabeled, sparse, and class-imbalanced. The central intuition is that relational semantics of concepts in label space are important for few-shot image recognition – we can borrow from familiar concepts to learn about new ones. In this paper, we show a formulation of this hypothesis by learning the graph distances between classes in the WordNet hierarchy.

From a scientific perspective, this research paper proposes a more “human-like” learning procedure that extends beyond few-shot learning and provides a formulation for meta-learning and active learning. From a medical application perspective, we can do better drug discovery over protein-protein interaction networks with less data. Overall, the diversity and importance of potential applications, combined with a burst of research interest in this field, suggests graph-based few-shot learning as a promising research direction.

REFERENCES

Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *Journal of machine learning research*, 6(Oct):1705–1749, 2005.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka, and Tom M Mitchell. Toward an architecture for never-ending language learning. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. Neil: Extracting visual knowledge from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1409–1416, 2013.

Guneet S Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. *arXiv preprint arXiv:1909.02729*, 2019.

Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1126–1135. JMLR. org, 2017.

Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks. *arXiv preprint arXiv:1711.04043*, 2017.

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 855–864. ACM, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 25*, pp. 1097–1105. Curran Associates, Inc., 2012. URL http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

Jure Leskovec and Rok Sosič. Snap: A general-purpose network analysis and graph-mining library. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(1):1, 2016.

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

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 701–710. ACM, 2014.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pp. 4077–4087, 2017.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pp. 935–943, 2013.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1199–1208, 2018.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning, 2016.

Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6857–6866, 2018.

Jason Weston, Frédéric Ratle, Hossein Mobahi, and Ronan Collobert. Deep learning via semi-supervised embedding. In *Neural Networks: Tricks of the Trade*, pp. 639–655. Springer, 2012.