

Graph convolutional networks for learning with few clean and many noisy labels

Ahmet Iscen¹, Giorgos Tolias², Yannis Avrithis³, Ondřej Chum², and Cordelia Schmid¹

¹ Google Research

² VRG, Faculty of Electrical Engineering, Czech Technical University in Prague

³ Inria, Univ Rennes, CNRS, IRISA

Abstract. In this work we consider the problem of learning a classifier from noisy labels when a few clean labeled examples are given. The structure of clean and noisy data is modeled by a graph per class and Graph Convolutional Networks (GCN) are used to predict class relevance of noisy examples. For each class, the GCN is treated as a binary classifier, which learns to discriminate clean from noisy examples using a weighted binary cross-entropy loss function. The GCN-inferred “clean” probability is then exploited as a relevance measure. Each noisy example is weighted by its relevance when learning a classifier for the end task. We evaluate our method on an extended version of a few-shot learning problem, where the few clean examples of novel classes are supplemented with additional noisy data. Experimental results show that our GCN-based cleaning process significantly improves the classification accuracy over not cleaning the noisy data, as well as standard few-shot classification where only few clean examples are used.

1 Introduction

State-of-the-art deep learning methods require a large amount of manually labeled data. The need for supervision may be reduced by decoupling representation learning from the end task and/or using additional training data that is unlabeled, weakly labeled (with noisy labels), or belong to different domains or classes. Example approaches are *transfer learning* [39], *unsupervised representation learning* [39], *semi-supervised learning* [42], *learning from noisy labels* [16] and *few-shot learning* [33].

However, for several classes, only very few or even no clean labeled examples might be available at the representation learning stage. *Few-shot learning* severely limits the number of labeled samples on the end task, while the representation is learned on a large training set of different classes [12,33,38]. Nevertheless, in many situations, more data with noisy labels can be acquired or is readily available for the end task.

One interesting mix of few-shot learning with additional large-scale data is the work of Douze *et al.* [5], where labels are propagated from few clean labeled examples to a large-scale collection. This collection is unlabeled and actually

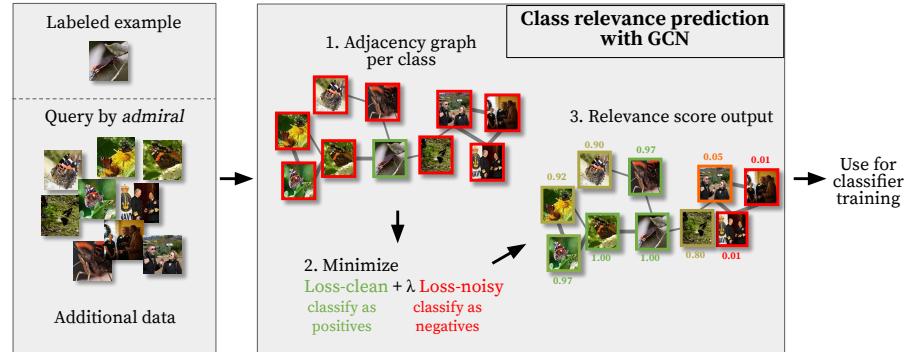


Fig. 1. Overview of our cleaning approach for 1-shot learning with noisy examples. We use the class name *admiral* to crawl noisy images from the web and create an adjacency graph based on visual similarity. We then assign a relevance score to each noisy example with a graph convolutional network (GCN). Relevance scores are displayed next to the images.

contains data for many more classes than the end task. Their method overall improves the classification accuracy, but at an additional computational cost. It is a *transductive* method, *i.e.*, instead of learning a parametric classifier, the large-scale collection is still necessary at inference.

In this work, we learn a classifier from a few clean labeled examples and additional weakly labeled data, while the representation is learned on different classes, similarly to few-shot learning. We assume that the class names are known, and we use them to search an existing large collection of images with textual description. The result is a set of images with novel class labels, but potentially incorrect (noisy). As shown in Figure 1, we clean this data using a *graph convolutional network* (GCN) [17], which learns to predict a class relevance score per image based on connections to clean images in the graph. Both the clean and the noisy images are then used to learn a classifier, where the noisy examples weighted by relevance. Unlike most existing work, our method operates independently per class and applies when clean labeled examples are few or even only one per class.

We make the following contributions:

1. We learn a classifier on a *large-scale weakly-labeled* collection jointly with only a *few clean labeled* examples.
2. To our knowledge, we are the first to use a GCN to clean noisy data: we cast a GCN as a *binary classifier* which learns to discriminate clean from noisy data, and we use its inferred “clean” probabilities as a relevance score per example.
3. We apply our method to two few-shot learning benchmarks and show significant improvement in accuracy, outperforming the method by Douze *et al.* [5] using the same large-scale collection of data without labels.

2 Related work

Learning with noisy labels is often concerned with estimating or learning a *transition matrix* [24,25,34] or *knowledge graph* [19] between labels and correcting the loss function, which does not apply in our case since the true classes in the noisy data is unknown. Most recent work on learning from large-scale weakly-labeled data focuses on learning the *representation* through *metric learning* [18,40], *bootstrapping* [28], or *distillation* [19]. In our case however, since the clean labeled examples are scarce, we need to keep the representation mostly fixed.

Dealing with the noise by thresholding [18], outlier detection [40] or *reweighting* [20], is applicable while the representation is learned, based on the gradient of the loss [30]. In contrast, the relatively-shallow GCN that we propose effectively decouples reweighting from both representation learning and classifier learning. *Learning to clean* the noisy labels [36] typically assumes adequate human verified labels for training, which again is not the case in this work.

Few-shot learning. *Meta-learning* [37] refers to learning at two levels, where generic knowledge is acquired before adapting to more specific tasks. In few-shot learning, this translates to learning on a set of *base classes* how to learn from few examples on a distinct set of *novel classes* without overfitting. For instance, *optimization meta-learning* [6,7,27] amounts to learning a model that is easy to fine-tune in few steps. In our work, we study an extension of few-shot learning where more novel class instances are available, reducing the risk of overfitting when fine-tuning the model. *Metric learning* approaches learn how to compare queries for instance to few examples [38] or to the corresponding *class prototypes* [33]. Hariharan and Girshick [12] and Wang *et al.* [41] learn how to generate novel-class examples, which is not needed when more data is actually available.

Gidaris and Komodakis [9] learn on base classes a simpler cosine similarity-based parametric classifier, or simply *cosine classifier*, without meta-learning. The same classifier has been introduced independently by Qi *et al.* [26], who further fine-tune the network, assuming access to the base class training set. A recent survey [3] confirms the superiority of the cosine classifier to previous work including meta-learning [6]. We use the cosine classifier in this work, both for base and novel classes.

Making use of *unlabeled data* has been little explored in few-shot learning until recently. Ren *et al.* [29] introduce a semi-supervised few-shot classification task, where some labels are unknown. Liu *et al.* [21] follow the same semi-supervised setup, but use graph-based *label propagation* (LP) [45] for classification and consider all test images jointly. These methods assume a meta-learning scenario, where only small-scale data is available at each training episode; arguably, such a small amount of data limits the representation adaptation and generalization to unseen data. Similarly, Rohrbach *et al.* [31] use label propagation in a *transductive* setting, but at a larger scale assuming that all examples come from a set of known classes. Douze *et al.* [5] extend to an even larger scale,

leveraging 100M unlabeled images in a graph without using additional text information. We focus on the latter large-scale scenario using the same 100M dataset. However, we filter by text to obtain *noisy data* and follow an *inductive* approach by training a classifier for novel classes, such that the 100M collection is not needed at inference.

Graph neural networks are generalizations of convolutional networks to non-Euclidean spaces [1]. Early *spectral methods* [2,14] have been succeeded by *Chebyshev polynomial* approximations [4], which avoid the high computational cost of computing eigenvectors. *Graph convolutional networks* (GCN) [17] provide a further simplification by a *first-order* approximation of graph filtering and are applied to *semi-supervised* [17] and subsequently *few-shot learning* [8]. Kipf and Welling [17] apply the loss function to labeled examples to make predictions on unlabeled ones. Similarly, Garcia and Bruna [8] use GCNs to make predictions on novel class examples. Gidaris and Komodakis [10] use Graph Neural Networks as denoising autoencoders to generate class weights for novel classes. By contrast, we cast GCNs as *binary classifiers* discriminating clean from noisy examples: we apply a loss function to all examples, and then use the inferred probabilities as a class relevance measure, effectively cleaning the data.

Our counter-intuitive objective of treating all noisy examples as negative can be compared to treating each example as a different class in *instance-level discrimination* [43]. In fact, our loss function is similar to *noise-contrastive estimation* (NCE) [11]. Our experiments show that our GCN-based classifier outperforms classical LP [45] used for a similar purpose by [31].

3 Problem formulation

We consider a space \mathcal{X} of examples. We are given a set $X_C \subset \mathcal{X}$ of examples, each having a *clean* (manually verified) label in a set C of classes with $|C| = K$. We assume that the number $|X_C^c|$ of examples⁴ labeled in each class $c \in C$ is only k , typically in $\{1, 2, 5, 10, 20\}$. We are also given an additional set X_N of examples, each with a set of *noisy* labels in C . The *extended* set of examples for class c is now $X_E^c = X_C^c \cup X_N^c$. Examples or sets of examples having clean (noisy) labels are referred to as clean (noisy) as well. The goal is to train a K -way classifier, using the additional noisy set in order to improve the accuracy compared to only using the small clean set.

We assume that we are given a feature extractor $g_\theta : \mathcal{X} \rightarrow \mathbb{R}^d$, mapping an example to a d -dimensional vector. For instance, when examples are images, the feature extractor is typically a *convolutional neural network* (CNN) and θ are the parameters of all layers.

In this work, we assume that the noisy set X_N is collected via web crawling. Examples are images accompanied by free-form text descriptions and/or user tags originating from community photo collections. To make use of text data, we assume that the names of the classes in C are given. An example in X_N is

⁴ For any set $X \subset \mathcal{X}$, we denote by X^c its subset of examples labeled in class $c \in C$.

given a label in class $c \in C$ if its textual information contains the name of class c ; it may then have none, one or more labels. In this way, we automatically infer labels for $X_{\mathcal{N}}$ without human effort, which are however *noisy*.

4 Cleaning with graph convolutional networks

We perform cleaning by predicting a *class relevance* measure for each noisy example in $X_{\mathcal{N}}^c$, independently for each class $c \in C$. To simplify the notation, we drop superscript c where possible in this subsection and we denote $X_{\mathcal{E}}^c$ by $\{x_1, \dots, x_k, x_{k+1}, \dots, x_N\}$, where $X_{\mathcal{C}}^c = \{x_1, \dots, x_k\}$ and $X_{\mathcal{N}}^c = \{x_{k+1}, \dots, x_N\}$. The features of these examples are similarly represented by matrix $V = [\mathbf{v}_1, \dots, \mathbf{v}_k, \mathbf{v}_{k+1}, \dots, \mathbf{v}_N] \in \mathbb{R}^{d \times N}$, where $\mathbf{v}_i = g_{\theta}(x_i)$ for $i = 1, \dots, N$.

We construct an affinity matrix $A \in \mathbb{R}^{N \times N}$ with elements $a_{ij} = [\mathbf{v}_i^\top \mathbf{v}_j]_+$ if examples \mathbf{v}_i and \mathbf{v}_j are reciprocal nearest neighbors in $X_{\mathcal{E}}^c$ and 0 otherwise. Matrix A has zero diagonal, but self-connections are added before A is normalized as $\tilde{A} = D^{-1}(A + I)$ with $D = \text{diag}((A + I)\mathbf{1})$ being the degree matrix of $A + I$ and $\mathbf{1}$ the all-ones vector.

Graph convolutional networks (GCNs) [17] are formed by a sequence of layers. Each layer is a function $f_{\Theta} : \mathbb{R}^{N \times N} \times \mathbb{R}^{l \times N} \rightarrow \mathbb{R}^{n \times N}$ of the form

$$f_{\Theta}(\tilde{A}, Z) = h(\Theta^\top Z \tilde{A}), \quad (1)$$

where $Z \in \mathbb{R}^{l \times N}$ represents the input features, $\Theta \in \mathbb{R}^{l \times n}$ holds the parameters of the layer to be learned, and h is a nonlinear activation function. Function f_{Θ} maps l -dimensional input features to n -dimensional output features.

In this work we consider a two-layer GCN with a scalar output per example. This network is a function $F_{\Theta} : \mathbb{R}^{N \times N} \times \mathbb{R}^{d \times N} \rightarrow \mathbb{R}^N$ given by

$$F_{\Theta}(\tilde{A}, V) = \sigma(\Theta_2^\top [\Theta_1^\top V \tilde{A}]_+ \tilde{A}), \quad (2)$$

where $\Theta = \{\Theta_1, \Theta_2\}$, $\Theta_1 \in \mathbb{R}^{d \times m}$, $\Theta_2 \in \mathbb{R}^{m \times 1}$, $[\cdot]_+$ is the positive part or ReLU function [23] and $\sigma(a) = (1 + e^{-a})^{-1}$ for $a \in \mathbb{R}$ is the sigmoid function. Function F_{Θ} performs feature propagation through the affinity matrix in an analogy to classical graph-based propagation methods for classification [45] or search [46].

The output $F_{\Theta}(\tilde{A}, V)$ is a vector of length N , with element $F_{\Theta}(\tilde{A}, V)_i$ in $[0, 1]$ representing a relevance value of example x_i for class c . To learn the parameters Θ , we treat the GCN as a *binary classifier* where target output 1 corresponds to clean examples and 0 to noisy. In particular, we minimize the loss function

$$L_{\mathcal{G}}(V, \tilde{A}; \Theta) = -\frac{1}{k} \sum_{i=1}^k \log(F_{\Theta}(\tilde{A}, V)_i) - \frac{\lambda}{N-k} \sum_{i=k+1}^N \log(1 - F_{\Theta}(\tilde{A}, V)_i). \quad (3)$$

This is a binary cross-entropy loss function where noisy examples are given an importance weight λ . Given the propagation on the nearest neighbor graph, and depending on the relative importance λ of the second term, noisy examples that are strongly connected to clean ones are still expected to receive high class

relevance, while noisy examples that are not relevant to the current class are expected to get a class relevance near zero.

The impact of parameter λ is validated in Section 6, where we show that the fewer the available clean images are (smaller k) the smaller the importance weight should be. As is standard practice for GCNs in classification [17], training is performed in batches of size N , that is the entire set of features.

Figure 2 shows examples of clean images, corresponding noisy ones and the predicted relevance. Using the visual similarity to the clean image, we can use relevance to resolve cases of polysemy, *e.g.* *black widow (spider)* vs. *black widow (superhero)*, or cases like *pineapple* vs. *pineapple juice*.

Discussion. Loss function (3) is similar to *noise-contrastive estimation* (NCE) [11] as used by We *et al.* [43] for *instance-level discrimination*, whereas we discriminate clean from noisy examples. The semi-supervised learning setup of GCNs [17] uses a loss function that applies only to the labeled examples, and makes discrete predictions on unlabeled examples. In our case, all examples contribute to the loss but with different importance, as we infer real-valued class relevance for the noisy examples, to be used for subsequent learning.

Function F_Θ in (2) reduces to a Multi-Layer Perceptron (MLP) when the affinity matrix A is zero, in which case all examples are disconnected. Using an MLP to perform cleaning would take each example into account independently of the others, while the GCN considers the collection of examples as a whole. MLP training is performed identically to GCN by minimizing (3). We compare the two alternatives in our experiments.

5 Learning a classifier with few clean and many noisy examples

Our cleaning process applies when the clean labeled examples are few, but assumes a feature extractor⁵ g_θ . That is, representation learning, label cleaning and classifier learning are decoupled. We perform GCN-based cleaning as described in Section 4, and learn a classifier by weighting examples according to class relevance. The process of training the classifier is described below.

5.1 Cosine-similarity based classifier

We use a *cosine-similarity based classifier* [9,26], or *cosine classifier* for short. Each class $c \in C$ is represented by a learnable parameter $\mathbf{w}_c \in \mathbb{R}^d$. The *prediction* of example $x \in \mathcal{X}$ is the class c of maximum cosine similarity $\hat{\mathbf{w}}_c^\top \hat{g}_\theta(x)$ ⁶

$$\pi_{\theta,W}(x) = \arg \max_c \hat{\mathbf{w}}_c^\top \hat{g}_\theta(x), \quad (4)$$

where $W = [\mathbf{w}_1, \dots, \mathbf{w}_K] \in \mathbb{R}^{d \times K}$.

⁵ For instance, after training on a different task or a set of classes other than C . Learning of the feature extractor used in our experiments is described in Appendix.

⁶ We denote the ℓ_2 -normalized counterpart of vector \mathbf{x} by $\hat{\mathbf{x}}$. Similarly, if $\mathbf{y} = f(x)$, we denote $\hat{\mathbf{y}}$ by $\hat{f}(x)$.

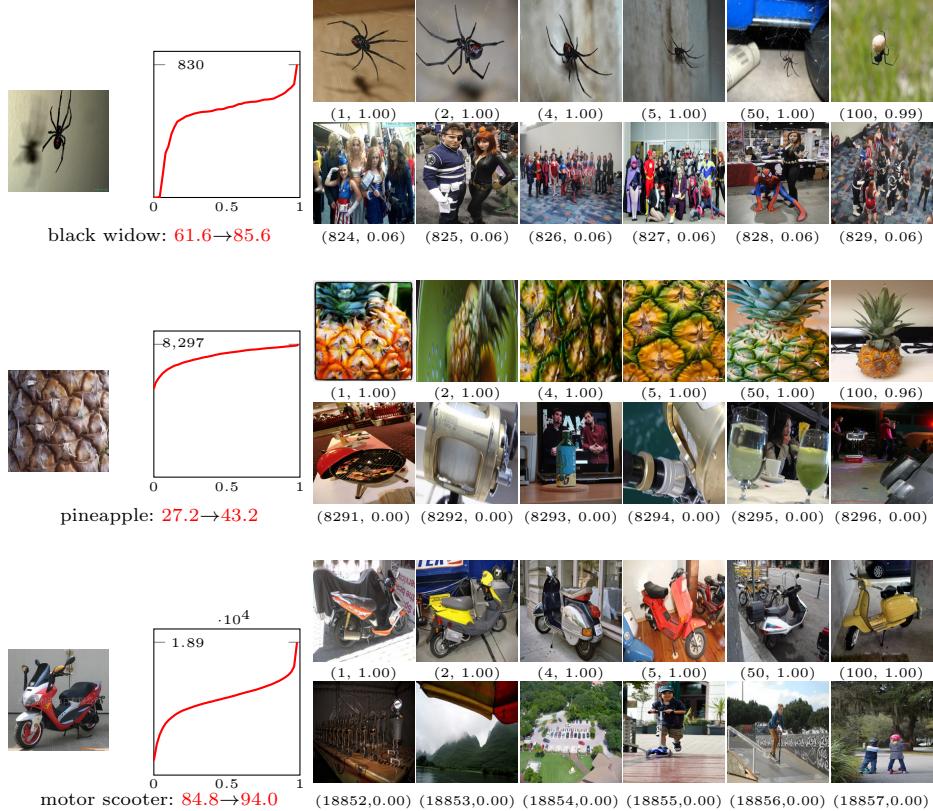


Fig. 2. Examples of clean images from the Low-Shot ImageNet Benchmark (left) for 1-shot classification, cumulative histogram of the predicted relevance for noisy images (middle) and representative noisy images (right) by descending order of relevance, with relevance value reported below. Test accuracy without and with additional data using class prototypes (5) is shown next to class names.

5.2 Classifier learning

The goal is to learn a K -way classifier for unseen data in \mathcal{X} . Unlike the typical few-shot learning task, each class contains a few clean and many noisy examples.

Prior to learning classifiers, training examples $x_i \in X_{\mathcal{E}}^c$ are weighted by their relevance $r(x_i)$ to class c . For a noisy example $x_i \in X_{\mathcal{N}}^c$, we define $r(x_i) = F_{\Theta}(\tilde{A}, V)_i$ where $F_{\Theta}(\tilde{A}, V)$ is the output vector of the GCN, while for a clean example $x_i \in X_{\mathcal{C}}^c$ we fix $r(x_i) = 1$. Note that optimizing (3) does not guarantee $F_{\Theta}(\tilde{A}, V)_i = 1$ for clean examples $x_i \in X_{\mathcal{C}}^c$. We define $r(X) = \sum_{x \in X} r(x)$ for any set $X \subset \mathcal{X}$.

We first assume that the given feature extractor is fixed and consider two different classifiers, namely class prototypes and cosine-similarity based classifier.

Then, this assumption is dropped and the classifier and feature representation are learned jointly by fine-tuning the entire network.

Class prototypes. For each class $c \in C$, we define *prototype* \mathbf{w}_c by

$$\mathbf{w}_c = \frac{1}{r(X_{\mathcal{E}}^c)} \sum_{x \in X_{\mathcal{E}}^c} r(x) g_{\theta}(x). \quad (5)$$

Prototypes are fixed vectors, not learnable parameters. Collecting them into matrix $W = [\mathbf{w}_1, \dots, \mathbf{w}_K] \in \mathbb{R}^{d \times K}$, K -way prediction is made by classifier $\pi_{\theta, W}$ (4).

Cosine classifier learning. Given examples $X_{\mathcal{E}}$, we learn a parametric cosine classifier with parameters $W = [\mathbf{w}_1, \dots, \mathbf{w}_K] \in \mathbb{R}^{d \times K}$ by minimizing the weighted cross entropy loss $L(C, X_{\mathcal{E}}, \theta; W)$ over W , given by

$$L(C, X_{\mathcal{E}}, \theta; W) = - \sum_{c \in C} \frac{1}{r(X_{\mathcal{E}}^c)} \sum_{x \in X_{\mathcal{E}}^c} r(x) \log(\sigma(s \hat{W}^{\top} \hat{g}_{\theta}(x))_c), \quad (6)$$

where $\sigma : \mathbb{R}^K \rightarrow \mathbb{R}^K$ is the softmax function with $\sigma(\mathbf{a})_c = e^{a_c} / \sum_{j \in C} e^{a_j}$ for $\mathbf{a} \in \mathbb{R}^K$, s is a *scale parameter* and $\hat{W} = [\hat{\mathbf{w}}_1, \dots, \hat{\mathbf{w}}_K]$. The parameters θ of the feature extractor are fixed. The scale parameter s is also fixed according to the training of the feature extractor. Prediction is made as in the previous case.

Deep network fine-tuning. An alternative is to drop the assumption that the feature extractor is fixed. In this case, we jointly learn the parameters θ of the feature extractor and W of the K -way cosine classifier by minimizing the right-hand side of (6). This requires access to examples $X_{\mathcal{E}}$, while for the previous two classifiers, access to features $g_{\theta}(x)$ is enough. Note that, due to over-fitting on the few available examples, such learning is typically avoided in a few-shot learning setup.

6 Experiments

6.1 Experimental setup

Datasets and task setup. We extend the *Low-Shot ImageNet benchmark* [12] by assuming many noisy examples in addition to the few clean ones. In this benchmark, the 1000 ImageNet classes [32] are split into 389 base classes and 611 novel classes. The validation set contains 193 base and 300 novel classes, and the test set the remaining 196 base and 311 novel classes. The base classes are used to learn the feature extractor (see supplementary material), while the novel classes form the set of classes C on which we apply the cleaning and learn the classifier. We only assume noisy examples for the novel classes, not for the base ones. Additionally, we apply a similar setup to the Places365 dataset [44]. We randomly choose 183 test and 182 validation classes. We use the model learned on the base classes of Low-Shot ImageNet benchmark as the feature extractor.

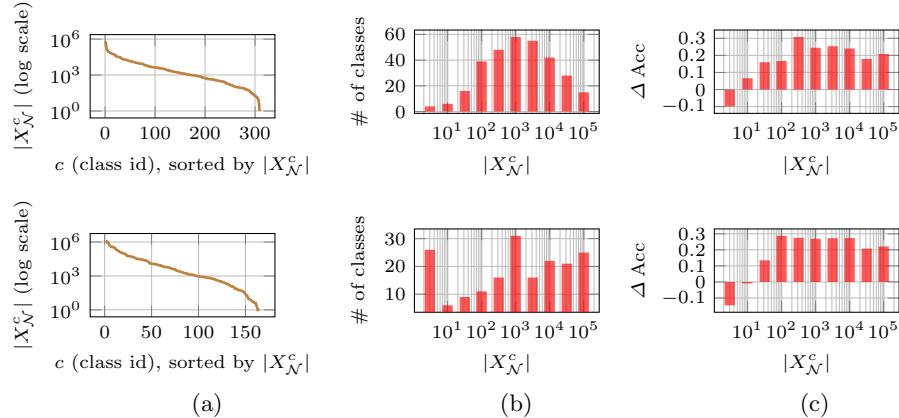


Fig. 3. Noisy data statistics for Low-Shot ImageNet (top) and Low-Shot Places365 (bottom). (a) Number of additional images collected from YFCC-100M per class c for all novel classes. (b) Number of classes per group, when groups are created according to $|X_N^c|$ in logarithmic scale. (c) Accuracy improvement ΔAcc (difference of accuracy between our method with noisy examples and the baseline without noisy examples) for prototype classifier, for same groups as in (b).

Therefore, all classes in Places365 dataset are considered *novel*. We refer to this setup as *Low-Shot Places365 benchmark*.

The standard benchmark includes k -shot classification, *i.e.* classification on k clean examples per class, with $k \in \{1, 2, 5, 10, 20\}$. We extend it to k clean and many noisy examples per class. Similar to the work of Hariharan and Girshick [12], we perform 5 episodes, each drawing a subset of k clean examples per class. We report the average top-5 accuracy over the 5 episodes on novel classes of the test set. Accuracy over all classes (base and novel) is reported in supplementary material.

Noisy data and statistics. We use the YFCC100M dataset [35] as a source of additional data with noisy labels. It contains approximatively 100M images collected from Flickr. Each image comes with a text description obtained from the user title and caption. We use the text description to obtain images with noisy labels, as discussed in Section 3.

Figure 3 (top) shows the statistics of noisy examples for Low-Shot ImageNet benchmark. The noisy examples for novel classes are long tailed in log scale (a). Noisy examples per class differ significantly for different classes, with a minimum of zero for classes *maillot* and *missile*, and a maximum of 620,142 for the class *church/church building*. There is a significant number of classes where we obtain less than 1000 extra examples, but we improve nevertheless; see Figure 3 (c). A small exception is 4 very rare classes out of 311, with around 3 additional images per class (leftmost bin in Figure 3 (b) and (c)). One could use more resources like web queries to collect additional data in real world applications.

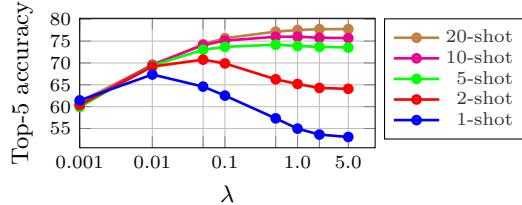


Fig. 4. Impact of λ on the validation set of the extended Low-shot ImageNet benchmark with YFCC-100M for noisy examples using class prototypes (5).

Figure 3 (bottom) presents the same statistics for Low-Shot Places365 benchmark. The trends are similar to those from Figure 3, except that there are more than 20 classes without any additional noisy data in this task (b). Nevertheless, there is a more consistent improvement in accuracy for classes that do have sufficient noisy data(c).

Representation and classifier learning. In most experiments, we use ResNet-10 [13] as feature extractor as in [9]. Classification for novel classes is performed with *class prototypes* (5), *cosine classifier learning* (6) or *deep network fine-tuning*. Hyper-parameters, such as batch size and number of epochs, are tuned on the validation set. We cross-validate the possible values of 512, 1024, 2048, 4096, and 8192 for batchsize and 10, 20, 30 and 50 for number of epochs. The learning rate starts from 0.1 and is reduced to 0.001 at the end of training with *cosine annealing* [22]. We handle the imbalance of the noisy set by normalizing by $r(X_c)$ in (6). Prototypes (5) are used to initialize the weights W of the cosine classifier in (6). We ignore examples x_i with relevance $r(x_i) < 0.1$ to reduce the complexity when fine-tuning the network. We also report results with ResNet-50 as feature extractor, using the model trained on base classes by [12]. Following [5], we apply PCA to the features to reduce their dimensionality to 256. Base classes are represented by class prototypes (5) in this case.

GCN training is performed with the Adam optimizer and a learning rate of 0.1 for 100 iterations. We use dropout with probability 0.5. The dimensionality of the input descriptors is $d = 512$ for ResNet-10 and $d = 256$ for ResNet-50 (after PCA). Dimensionality of the internal representation in (1) is $m = 16$. The affinity matrix is constructed with reciprocal top-50 nearest neighbors.

Baseline methods. We implement and evaluate several baseline methods:

1. *β -cleaning* assigns $r(x_i) = \beta$ to all additional examples. We report results for $\beta = 1.0$ (unit relevance score) and β^* , the optimal β for all k obtained on the validation set.
2. *Similarity* uses the scaled cosine similarity as the relevance weight, *i.e.* $r(x_i) = (1 + \mathbf{v}_i^\top \mathbf{x})/2$, where \mathbf{x} is the class prototype created with the features of clean examples.
3. *Linear* learns a linear binary classifier where the positive instances are the k labeled examples, and the negative examples are chosen randomly from other classes.

Method	$k=1$	2	5	10	20
FEW CLEAN EXAMPLES					
Class proto. [9]	45.3 \pm 0.65	57.1 \pm 0.37	69.3 \pm 0.32	74.8 \pm 0.20	77.8 \pm 0.24
FEW CLEAN & MANY NOISY EXAMPLES					
Similarity	49.8 \pm 0.29	56.3 \pm 0.27	64.2 \pm 0.32	68.4 \pm 0.14	71.2 \pm 0.12
β -weighting, $\beta = 1$	56.1 \pm 0.06	56.4 \pm 0.08	57.1 \pm 0.05	57.7 \pm 0.08	58.7 \pm 0.06
β -weighting, β^*	55.6 \pm 0.24	58.3 \pm 0.14	63.4 \pm 0.25	67.5 \pm 0.34	71.0 \pm 0.22
Linear	59.8 \pm 0.00	59.3 \pm 0.00	58.4 \pm 0.00	58.6 \pm 0.00	59.4 \pm 0.00
Label Propagation	62.6 \pm 0.35	67.0 \pm 0.41	74.6 \pm 0.30	76.3 \pm 0.23	77.7 \pm 0.18
MLP	63.6 \pm 0.41	68.8 \pm 0.42	73.7 \pm 0.25	75.6 \pm 0.21	77.6 \pm 0.21
Ours	67.8 \pm 0.10	70.9 \pm 0.30	73.9 \pm 0.17	76.1 \pm 0.12	78.2 \pm 0.14

Table 1. Comparison with baselines using noisy examples on the Low-shot ImageNet benchmark. We report top-5 accuracy on novel classes with classification by class prototypes (5).

- 4. *Label Propagation* (LP) [45] propagates information by a linear operation. It solves the linear system $(I - \alpha D^{-1/2} A D^{-1/2}) \mathbf{r}_c = \mathbf{y}_c$ [15] for each class c , where D is the degree matrix of A , $\alpha = 0.9$ and $\mathbf{y}_c \in \mathbb{R}^N$ is a k -hot binary vector indicating the clean (labeled) examples of class c . Relevance $r(x_i)$ is the i -th element $(\mathbf{r}_c)_i$ of the solution.
- 5. *MLP*, discussed in Section 4, learns a nonlinear mapping to assign relevance weights, but does not propagate over the graph. It is trained using (3) and therefore includes part of our contribution.

6.2 Experimental results

The impact of the importance weight λ is measured on the validation set and the best performing value is used on the test set for each value of k . Results on Low-shot ImageNet benchmark are shown in Figure 4. The larger the value of λ , the more the loss encourages noisy examples to be classified as negatives. As a consequence, large (small) λ results in smaller (larger) relevance, on average, for noisy examples. The optimal λ per value of k suggests that as the number of clean examples decreases, the need for additional noisy ones increases.

Comparison with baselines using additional data on Low-shot Imagenet benchmark is presented in Table 1. Qualitative results are presented in Figure 2. The use of additional data is mostly harmful for β -weighting except for 1 and 2-shot. MLP offers improvements in most cases, implying that it manages to appropriately down-weight irrelevant examples. The consistent improvement of GCN compared to MLP, especially large for small k , suggests that it is beneficial to incorporate relations, with the affinity matrix A modeling the structure of the feature space. LP is a classic approach that also uses A but is a linear operation with no parameters, and is inferior to our method. The gain of cleaning ($\beta = 1$ vs. ours) ranges from 11% to 20%.

Comparison with the state of the art on Low-shot Imagenet benchmark is presented in Table 2. We significantly improve the performance by using additional data and cleaning compared to a number of different approaches, including

METHOD	TOP-5 ACCURACY ON NOVEL CLASSES				
	$k=1$	2	5	10	20
RESNET-10 – FEW CLEAN EXAMPLES					
Proto.-Nets [33] [†]	39.3	54.4	66.3	71.2	73.9
Logistic reg. w/ H [41] [†]	40.7	50.8	62.0	69.3	76.5
PMN w/ H [41] [†]	45.8	57.8	69.0	74.3	77.4
Class proto. [9]	45.3±0.65	57.1±0.37	69.3±0.32	74.8±0.20	77.8±0.24
Class proto. w/ Att. [9]	45.8±0.74	57.4±0.38	69.6±0.27	75.0±0.29	78.2±0.23
RESNET-10 – FEW CLEAN & MANY NOISY EXAMPLES					
Ours - class proto. (5)	67.8±0.10	70.9±0.30	73.7±0.20	76.1±0.16	78.2±0.14
Ours - cosine (6)	73.2±0.14	75.3±0.25	75.6±0.24	78.5±0.32	80.7±0.26
Ours - fine-tune	74.1±0.19	76.2±0.28	77.7±0.23	80.6±0.31	82.6±0.24
RESNET-50 – FEW CLEAN EXAMPLES					
Proto.-Nets [33] [†]	49.6	64.0	74.4	78.1	80.0
PMN w/ H [41] [†]	54.7	66.8	77.4	81.4	83.8
RESNET-50 – FEW CLEAN & MANY UNLABELED EXAMPLES					
Diffusion [5] [†]	63.6±0.61	69.5±0.60	75.2±0.40	78.5±0.34	80.8±0.18
Diffusion - logistic [5] [†]	64.0±0.70	71.1±0.82	79.7±0.38	83.9±0.10	86.3±0.17
RESNET-50 – FEW CLEAN & MANY NOISY EXAMPLES					
Ours - class proto. (5)	69.7±0.44	73.7±0.56	77.0±0.20	79.9±0.30	81.9±0.29
Ours - cosine (6)	78.0±0.38	80.2±0.33	80.9±0.17	83.7±0.19	85.7±0.11
Ours - fine-tune	80.2±0.33	82.6±0.14	83.3±0.26	85.9±0.22	88.3±0.21

Table 2. Comparison to the state of the art on the Low-shot ImageNet benchmark. We report top-5 accuracy on novel classes. We use class prototypes (5), cosine classifier learning (6) and deep network fine-tuning for classification with our GCN-based data addition method. [†] denotes numbers taken from the corresponding papers. All other experiments are re-implemented by us.

the work by Gidaris and Komodakis [9], which is our starting point. As expected, the gain is more pronounced for small k , reaching more than 20% improvement for 1-shot novel accuracy.

Closest to ours is the work by Douze *et al.* [5], who use the same experimental setup and the same additional data, but without filtering by text or using noisy labels. We outperform this approach in all cases, while requiring much less computation: *offline*, we construct a separate small graph per class rather than a single graph over the entire 100M collection; *online*, we perform inference by cosine similarity to one prototype per class or a learned classifier rather than iterative diffusion on the entire collection. By ignoring examples that are not given any noisy label, we are only using a tiny fraction of the 100M collection: in particular, only 3,744,994 images for the 311-class test split of the Low-shot ImageNet benchmark. In contrast to [5], additional data brings improvement even at 20-shot with classifier learning or network fine-tuning. Most importantly, our approach does not require the entire 100M collection at inference.

Analysis of relevance weights. We manually label all the noisy examples from 20 classes in order to quantitatively measure the accuracy of the assigned relevance. We measure the noise ratio per class, *i.e.* the ratio of irrelevant (negative) noisy images to all noisy (positive and negative) images. Positive and negative

Method	$k=1$	2	5	10	20
FEW CLEAN EXAMPLES					
Class proto. [9]	28.7 \pm 1.12	38.0 \pm 0.37	50.5 \pm 0.51	57.9 \pm 0.35	62.3 \pm 0.25
FEW CLEAN & MANY NOISY EXAMPLES - CLASS PROTO. (5)					
β -weighting, $\beta = 1$	44.0 \pm 0.34	45.7 \pm 0.22	48.4 \pm 0.31	50.0 \pm 0.12	50.8 \pm 0.25
Label Propagation	39.6 \pm 0.78	46.5 \pm 0.22	54.8 \pm 0.42	59.6 \pm 0.11	62.0 \pm 0.14
MLP	46.9 \pm 0.78	50.1 \pm 0.38	55.4 \pm 0.29	59.2 \pm 0.26	61.5 \pm 0.31
Ours	47.1 \pm 0.70	50.5 \pm 0.31	55.1 \pm 0.50	59.0 \pm 0.32	61.9 \pm 0.22
FEW CLEAN & MANY NOISY EXAMPLES - OTHER CLASSIFIERS					
Ours - cosine (6)	50.7 \pm 0.61	53.5 \pm 0.49	57.0 \pm 0.54	59.8 \pm 0.22	62.3 \pm 0.12
Ours - fine-tune	51.8 \pm 0.69	54.8 \pm 0.57	59.5 \pm 0.63	62.9 \pm 0.39	66.0 \pm 0.27

Table 3. Comparison with baselines using noisy examples on the Low-shot Places365 benchmark. We report top-5 accuracy on novel classes.

images are defined according to the manual labels. The 20 classes are selected such that 10 of them have the highest 1-shot accuracy, and the rest have the lowest. This allows us to examine success and failure cases.

In the case of 1-shot classification ($k = 1$), the average relevance weight is 0.71 for positive and 0.40 for negative examples. A success case is the “bee eater” class with noise ratio equal to 0.52. Our method achieves 98.4% accuracy for 1-shot classification, compared to 68.8% without any additional data. The average relevance weight is 0.99 for positive examples and 0.25 for negative examples of this class. One failure case is the “muzzle” class; it corresponds to the muzzle of an animal. The noise ratio is high; 94% of the 980 collected images are not relevant with most being animals without a muzzle or a firearm. The 1-shot classification accuracy without noisy data is 4%. Our method offers only a small increase to 8%. This can be explained by inaccurate relevance weights, which are on average 0.18 for positive and 0.30 for negative examples.

Experiments in Low-Shot Places365 are reported in Table 3. Our results indicate that our method consistently outperforms the baselines on this benchmark as well. Note that *MLP*, which is also competitive for this task, is trained with our proposed loss function 3. This is our contribution as well as the use of GCN. These methods significantly improve over existing methods, such as Label Propagation [45]. Further improvements are brought by cosine classifier learning (6) and deep network fine-tuning.

We also present qualitative results on Low-Shot Places365 in Figure 5. The first example at the top shows that top-ranked images by relevance depict different views of cafeterias for the *cafeteria* class, while bottom-ranked images depict food served in a cafeteria, which are irrelevant to our task. Similarly, our method assigns high relevance to images of soccer stadiums and low relevance to soccer players for the *soccer* class. Finally, our method finds similar images to the clean example for the *ruin* class. In general, top-ranking images exhibit diversity and are not just near-duplicates.

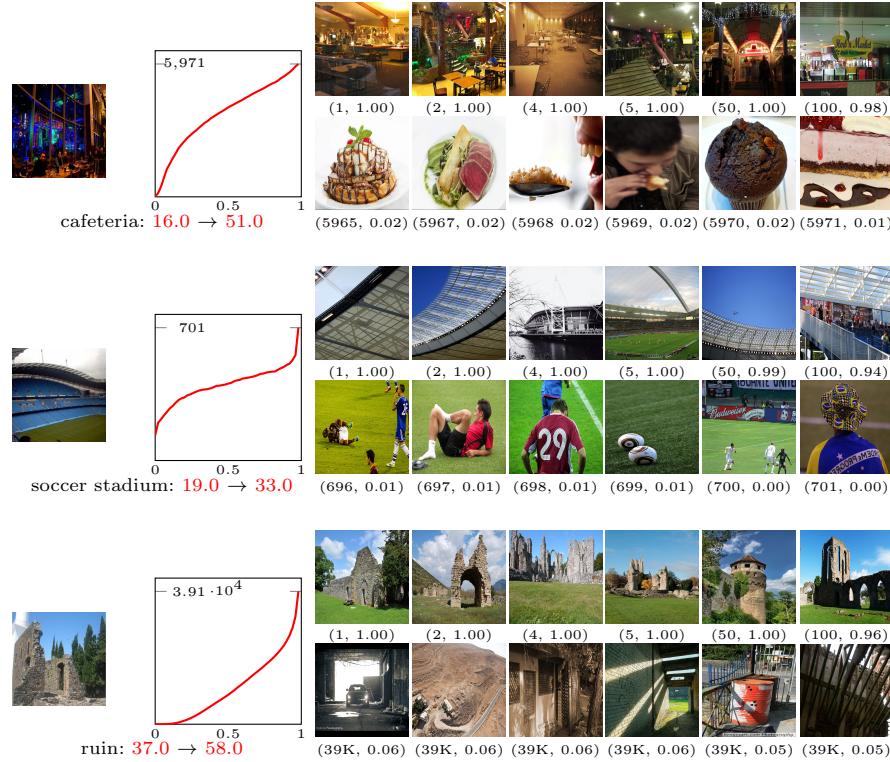


Fig. 5. Examples of clean images on Low-Shot Places365 Benchmark (left) for 1-shot classification, cumulative histogram of the predicted relevance for noisy images (middle), and representative noisy images (right), each having its position in the descending ranked list according to relevance value reported below. Test accuracy without and with additional data using prototypes (5) is shown next to class names.

7 Conclusions

In this paper we have introduced a new method for assigning class relevance to noisy images obtained by textual queries with class names. Our approach leverages one or a few labeled images per class and relies on a graph convolutional network (GCN) to propagate visual information from the labeled images to the noisy ones. The GCN is trained as a binary classifier discriminating clean from noisy examples using a weighted binary cross-entropy loss function and inferring “clean” probability as a relevance score for that class. Experimental results show that using noisy images weighted by this relevance score significantly improves the classification accuracy.

Acknowledgements. This work is funded by MSMT LL1901 ERC-CZ grant and OP VVV funded project CZ.02.1.01/0.0/0.0/16_019/0000765 “Research Center for Informatics”.

References

1. Bronstein, M.M., Bruna, J., LeCun, Y., Szlam, A., Vandergheynst, P.: Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine* **34**(4), 18–42 (2017)
2. Bruna, J., Zaremba, W., Szlam, A., Lecun, Y.: Spectral networks and locally connected networks on graphs. In: ICLR (2014)
3. Chen, W.Y., Liu, Y.C., Kira, Z., Wang, Y.C.F., Huang, J.B.: A closer look at few-shot classification. ICLR (2019)
4. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. In: NeurIPS (2016)
5. Douze, M., Szlam, A., Hariharan, B., Jégou, H.: Low-shot learning with large-scale diffusion. In: CVPR (2018)
6. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: ICML (2017)
7. Finn, C., Xu, K., Levine, S.: Probabilistic model-agnostic meta-learning. In: NeurIPS (2018)
8. Garcia, V., Bruna, J.: Few-shot learning with graph neural networks. In: ICLR (2018)
9. Gidaris, S., Komodakis, N.: Dynamic few-shot visual learning without forgetting. In: CVPR (2018)
10. Gidaris, S., Komodakis, N.: Generating classification weights with gnn denoising autoencoders for few-shot learning. In: CVPR (2019)
11. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In: International Conference on Artificial Intelligence and Statistics (2010)
12. Hariharan, B., Girshick, R.: Low-shot visual recognition by shrinking and hallucinating features. In: CVPR (2017)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
14. Henaff, M., Bruna, J., LeCun, Y.: Deep convolutional networks on graph-structured data. arXiv preprint arXiv:1506.05163 (2015)
15. Iscen, A., Tolias, G., Avrithis, Y., Furón, T., Chum, O.: Efficient diffusion on region manifolds: Recovering small objects with compact cnn representations. In: CVPR (2017)
16. Joulin, A., van der Maaten, L., Jabri, A., Vasilache, N.: Learning visual features from large weakly supervised data. In: ECCV (2016)
17. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: ICLR (2017)
18. Lee, K.H., He, X., Zhang, L., Yang, L.: CleanNet: Transfer learning for scalable image classifier training with label noise. In: CVPR (2018)
19. Li, Y., Yang, J., Song, Y., Cao, L., Luo, J., Li, L.J.: Learning from noisy labels with distillation. In: ICCV (2017)
20. Liu, T., Tao, D.: Classification with noisy labels by importance reweighting. *IEEE Trans. PAMI* **38**(3) (2015)
21. Liu, Y., Lee, J., Park, M., Kim, S., Yang, E., Hwang, S.J., Yang, Y.: Learning to propagate labels: Transductive propagation network for few-shot learning. In: ICLR (2019)
22. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. ICLR (2017)

23. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: ICML (2010)
24. Natarajan, N., Dhillon, I.S., Ravikumar, P.K., Tewari, A.: Learning with noisy labels. In: NeurIPS (2013)
25. Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., Qu, L.: Making deep neural networks robust to label noise: A loss correction approach. In: CVPR (2017)
26. Qi, H., Brown, M., Lowe, D.G.: Low-shot learning with imprinted weights. In: CVPR (June 2018)
27. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning. In: ICLR (2017)
28. Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., Rabinovich, A.: Training deep neural networks on noisy labels with bootstrapping. ICLR (2015)
29. Ren, M., Ravi, S., Triantafillou, E., Snell, J., Swersky, K., Tenenbaum, J.B., Larochelle, H., Zemel, R.S.: Meta-learning for semi-supervised few-shot classification. In: ICLR (2018)
30. Ren, M., Zeng, W., Yang, B., Urtasun, R.: Learning to reweight examples for robust deep learning. In: ICML (2018)
31. Rohrbach, M., Ebert, S., Schiele, B.: Transfer learning in a transductive setting. In: NeurIPS (2013)
32. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. IJCV **115**(3) (2015)
33. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: NeurIPS (2017)
34. Sukhbaatar, S., Bruna, J., Paluri, M., Bourdev, L., Fergus, R.: Training convolutional networks with noisy labels. arXiv preprint arXiv:1406.2080 (2014)
35. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Commun. ACM **59**(2) (2016)
36. Veit, A., Alldrin, N., Chechik, G., Krasin, I., Gupta, A., Belongie, S.: Learning from noisy large-scale datasets with minimal supervision. In: CVPR (July 2017)
37. Vilalta, R., Drissi, Y.: A perspective view and survey of meta-learning. Artificial intelligence review **18**(2), 77–95 (2002)
38. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. In: NeurIPS (2016)
39. Wang, X., Gupta, A.: Unsupervised learning of visual representations using videos. In: CVPR (2015)
40. Wang, Y., Liu, W., Ma, X., Bailey, J., Zha, H., Song, L., Xia, S.T.: Iterative learning with open-set noisy labels. In: CVPR (2018)
41. Wang, Y.X., Girshick, R., Hebert, M., Hariharan, B.: Low-shot learning from imaginary data. In: CVPR (2018)
42. Weston, J., Ratle, F., Collobert, R.: Deep learning via semi-supervised embedding. In: ICML (2008)
43. Wu, Z., Xiong, Y., Yu, S., Lin, D.: Unsupervised feature learning via non-parametric instance-level discrimination. CVPR (2018)
44. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Trans. PAMI **40**(6) (2017)
45. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B.: Learning with local and global consistency. In: NeurIPS (2003)
46. Zhou, D., Weston, J., Gretton, A., Bousquet, O., Schölkopf, B.: Ranking on data manifolds. In: NeurIPS (2003)

Supplementary Material: Graph convolutional networks for learning with few clean and many noisy labels

Anonymous ECCV submission

Paper ID 1060

A The role of base classes

The proposed method is applicable with any given feature extractor. Herein, we describe the learning of the feature extractor on a set of base classes according to a standard few-shot learning setup and benchmark [12]. Then, we describe the extended classifiers to the union of all classes, *i.e.* base classes and novel classes, which are the ones used in Section 5.

A.1 Representation learning on base classes

We are given a set $X_B \subset \mathcal{X}$ of examples, each having a clean label in a set of *base classes* C_B with $|C_B| = K_B$. Base classes C_B are disjoint from C , which are also known as novel classes. These data are used to learn a feature representation, *i.e.* a feature extractor g_θ , by learning a K_B -way base-class classifier for unseen data in \mathcal{X} . The parameters θ of the feature extractor and W_B of the classifier are jointly learned by minimizing the cross entropy loss

$$L_B(C_B, X_B; \theta, W_B) = - \sum_{c \in C_B} \frac{1}{|X_B^c|} \sum_{x \in X_B^c} \log(\sigma(s \hat{W}_B^\top \hat{g}_\theta(x))_c). \quad (1)$$

The learned feature extractor parameters θ and the learned scale parameter s are used by our method as described Sections 4 and 5.

A.2 Classification on all classes

The classifier parameters W_B are used, combined with classifier parameters W learned as described in Section 5, for classification on *all classes* $C_A = C \cup C_B$.

Class prototypes. The concatenated parameter matrix $W_A = [W_B, W]$ is used for K_A -way prediction on all (base and novel) classes by π_{θ, W_A} , where $K_A = K + K_B$. W_B is learned according to $L_B(C_B, X_B; \theta, W_B)$ (1), while W is learned according to (5).

Cosine classifier learning. Prediction on all classes is made as in the previous case, but W is learned according to (6).

Deep network fine-tuning. We now assume that base class examples are accessible too and, given all examples $X_A = X_B \cup X_E$, we jointly learn the parameters θ of the feature extractor and $W_A = [W_B, W]$ of the K_A -way cosine classifier for all classes by minimizing loss function

$$L_A(C_A, X_A; \theta, W_A) = L_B(C_B, X_B; \theta, W_B) + L(C, X_E; \theta, W). \quad (2)$$

METHOD	TOP-5 ACCURACY ON ALL CLASSES				
	$k=1$	2	5	10	20
RESNET-10 – FEW CLEAN EXAMPLES					
Proto.-Nets [33] [†]	49.5	61.0	69.7	72.9	74.6
Logistic reg. w/ H [41] [†]	54.4	61.0	69.0	73.7	76.5
PMN w/ H [41] [†]	40.8	49.9	64.2	71.9	76.9
Class proto. [9]	57.0±0.36	64.7±0.16	72.5±0.18	75.8±0.16	77.4±0.19
Class proto. w/ Att. [9]	58.1±0.48	65.2±0.15	72.9±0.25	76.6±0.18	78.8±0.16
RESNET-10 – FEW CLEAN & MANY NOISY EXAMPLES					
Ours - class proto. (5)	70.3±0.05	72.1±0.18	74.1±0.12	75.6±0.13	76.9±0.09
Ours - cosine (6)	72.4±0.07	73.4±0.21	77.2±0.20	78.8±0.21	79.2±0.17
Ours - fine-tune	76.0±0.10	77.3±0.13	78.7±0.19	80.7±0.25	82.2±0.14
RESNET-50 – FEW CLEAN EXAMPLES					
Proto.-Nets [33] [†]	61.4	71.4	78.0	80.0	81.1
PMN w/ H [41] [†]	65.7	73.5	80.2	82.8	84.5
RESNET-50 – FEW CLEAN & MANY NOISY EXAMPLES					
Ours - class proto. (5)	73.8±0.33	76.6±0.36	78.9±0.19	80.8±0.21	82.2±0.14
Ours - cosine (6)	78.2±0.25	79.6±0.23	80.4±0.18	82.4±0.19	84.1±0.09
Ours - fine-tune	81.6±0.20	83.2±0.16	84.3±0.23	86.2±0.17	87.8±0.03

Table 1. Comparison to the state of the art on the Low-shot ImageNet benchmark. We report top-5 accuracy on all classes. We use class prototypes (5), cosine classifier learning (6) and deep network fine-tuning for classification with our GCN-based data addition method. [†] denotes numbers taken from the corresponding papers. All other experiments are re-implemented by us.

Note that in contrast to (6), the last term of (2) optimizes parameters θ too. As mentioned earlier, such learning is typically avoided in a few-shot learning setup. In few cases, it takes the form of fine-tuning including all base class data [26], or only lasts for a few iterations when the base class data is not accessible [6].

A.3 Results on all classes

We report the accuracy over all classes in Table 1. When fine-tuning the network by (2), the learned W is used to initialize the corresponding part of W_A and we train all layers for 10 epochs with learning rate 0.01. The results indicate that our method still brings significant improvements when all classes are used.

B Results on Mini-Imagenet

We evaluate the proposed method on another popular benchmark, *i.e.* few-shot learning on Mini-ImageNet [38]. The dataset is a subset of ImageNet [32], and contains 100 different classes, split into 64 base, 16 validation and 20 test classes [27]. Each class contains 600 images that are re-sized to a resolution of 84×84 . We use the ConvNet-128 model with cosine classifier, following [9]. Novel categories are classified using class prototypes (5).

Method	$k=1$	$k=5$
FEW CLEAN EXAMPLES		
Class proto. [9]	54.2 ± 0.77	71.2 ± 0.61
Class proto. w/ Att. [9]	56.2 ± 0.81	72.9 ± 0.62
FEW CLEAN & MANY NOISY EXAMPLES - CLASS PROTO. (5)		
β -weighting, $\beta = 1$	63.5 ± 0.77	65.2 ± 0.81
Label Propagation	67.0 ± 0.74	74.8 ± 0.61
MLP	65.9 ± 0.78	73.9 ± 0.63
Ours	68.2 ± 0.76	74.7 ± 0.59

Table 2. Comparison with baselines using noisy examples on the Mini-ImageNet dataset. We report the accuracy for 5-way k -shot experiments where $k = 1$ and $k = 5$.

Table 2 shows the accuracy on Mini-Imagenet for the 5-way k -shot classification scenario with $k = 1$ and $k = 5$. We report the average accuracy over 600 trials along with the confidence interval. Our method brings significant improvements for $k = 1$, showing its generalization across different few-shot datasets and benchmarks.