

Weakly Supervised Image Classification through Noise Regularization

Mengying Hu^{1,3}, Hu Han^{1,2}, Shiguang Shan^{1,2,3,4}, Xilin Chen^{1,3}

¹ Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China

² Peng Cheng Laboratory, Shenzhen, China

³ University of Chinese Academy of Sciences, Beijing 100049, China

⁴ CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai, China

mengying.hu@ict.ac.cn, {hanhu, sgshan, xlchen}@ict.ac.cn

Abstract

Weakly supervised learning is an essential problem in computer vision tasks, such as image classification, object recognition, etc., because it is expected to work in the scenarios where a large dataset with clean labels is not available. While there are a number of studies on weakly supervised image classification, they usually limited to either single-label or multi-label scenarios. In this work, we propose an effective approach for weakly supervised image classification utilizing massive noisy labeled data with only a small set of clean labels (e.g., 5%). The proposed approach consists of a clean net and a residual net, which aim to learn a mapping from feature space to clean label space and a residual mapping from feature space to the residual between clean labels and noisy labels, respectively, in a multi-task learning manner. Thus, the residual net works as a regularization term to improve the clean net training. We evaluate the proposed approach on two multi-label datasets (OpenImage and MS COCO2014) and a single-label dataset (Clothing1M). Experimental results show that the proposed approach outperforms the state-of-the-art methods, and generalizes well to both single-label and multi-label scenarios.

1. Introduction

Weakly supervised learning is receiving increasing attention in many computer vision tasks, e.g., image classification, object recognition, etc., because incomplete and inaccurate annotations widely exist in many practical scenarios. For example, the diverse knowledge levels of different subjects can lead to different understanding of the same classes of images. In addition, a massive dataset may be collected automatically, and annotated by pre-trained models to re-

duce the cost in annotation time and expense, and only a small set of the labels can be verified by humans. It is challenging for the traditional supervised learning methods to work on such datasets with noisy labels. Therefore, weakly supervised learning from noisy data becomes valuable for practical applications and has drawn increasing attentions in recent years [5, 10, 11, 14, 25, 30].

Existing weakly supervised learning methods on image classification usually have certain assumptions with the noise label type, i.e., single-label noise or multi-label noise. Patrini et al. [23] defined a matrix T to describe the flipping relation between each two classes under the single label assumption. Veit et al. [30] implicitly learned a structure in the label space to do the prediction of multi-label. Both assumptions have their own characteristics. Single-label noise can introduce methods like clustering similar images in the training process [14] while multi-label noise can use label-

duce the cost in annotation time and expense, and only a small set of the labels can be verified by humans. It is challenging for the traditional supervised learning methods to work on such datasets with noisy labels. Therefore, weakly supervised learning from noisy data becomes valuable for practical applications and has drawn increasing attentions in recent years [5, 10, 11, 14, 25, 30].

Corresponding author.

to-label relation to make the algorithm more robust [33]. Although these assumptions help to improve the model performance, it limits the generalization ability of model from single-label dataset to multi-label dataset. For single-label noise learning methods [14, 23, 34], they cannot be applied to the multi-label data due to their strict assumption. For multi-label noise learning methods [20, 30, 33], their effectiveness on the single-label data is unknown.

In this work, we focus on reconciling the gap between single-label and multi-label in weakly supervised image classification. In our observation, although previous methods use different assumptions to assist in classifier learning, the key idea of them is to distinguish between the reliable and unreliable parts in the massive noisy labels, e.g., learning a mapping from noisy label space to clean label space or filtering out some noisy labels by utilizing the characteristic of noise. As shown in Fig. 1, for methods which use assumptions of label-to-label or image-to-label relationship, the reliable and unreliable parts of the noisy labels are determined by the strength of positive or negative correlations.

In this paper, we propose a weakly supervised learning approach for image classification that can automatically identify the reliable labels from the massive noisy labels. We expect that the proposed method can leverage massive noisy labeled data with a small set of clean labels to obtain a more robust image classification model. In addition, we expect that the proposed approach can generalize to both single-label and multi-label image classification. An overview of our approach is shown in Fig. 2. The proposed approach consists of a stem net (i.e., ResNet-50), a clean net, and a residual net. The stem net is used for shared feature learning. The clean net and the residual net are responsible for learning a mapping from feature space to clean label space and a residual mapping from feature space to the residual between clean labels and noisy labels, respectively. Similar to [11, 30], we only use a small fraction of dataset with clean labels to assist in the network training. We supervise the clean net using only the clean labels, and supervise the sum of clean net and residual net using noisy labels, respectively. The residual net works as a regularization term to improve the clean net training so that it can utilize the reliable information among the massive noisy data, while avoiding big influence by the unreliable information. Experimental results show that the proposed approach outperforms the state-of-the-art methods, and generalizes well to both single-label and multi-label scenarios.

Our contributions are as follows. (i) Our approach models the noisy label distribution via a residual net, and uses it to regularize the training of clean net so that the clean net can leverage the reliable part in the massive noisy data to improve the classification performance. (ii) Our approach has good generalization ability, and can work for both single-label and multi-label image classification tasks.

2. Related Work

Most of the weakly supervised image classification methods aim to learn from data with only noisy labels. One category of methods aims to distinguish the noise data from the entire dataset [2, 14, 17, 18, 24, 28, 32]. These methods usually focus on finding the difference between noisy data and clean data. Brodley and Friedl [2] used a set of classifiers to do outliers removal before training, under the assumption that outliers are likely noisy data. Reed et al. [24] used a bootstrap technique which could dynamically update the supervision under a prediction consistency assumption to filter out the potential noisy labels. Wang et al. [32] used contractive loss to distinguish between data with noisy labels and data with clean labels. Lee et al. [14] introduced a reference set to get a representative of the noisy set and used it for noise detection and image classification.

Another category of weakly supervised learning methods is to explore the new design of loss function [1, 6, 19, 21, 23] or network architecture [20, 22, 26] to achieve noise robust learning. These methods do not aim to separate noisy labels from all the labels explicitly. Instead, they incorporated the label transition process into the classification network design and aimed at learning a robust end-to-end classification model. Sukhbaatar et al. [26] used a single linear layer to model the label transition process from clean labels to noisy labels, while Patrini et al. [23] replaced the layer with estimating a noise transition matrix. Misra et al. [20] proposed an image-based probability model on multi-label noise to describe the relation between visually present (clean) labels and human-centric (noisy) labels.

Our method belongs to another stream, which under the assumption that few clean labels are known [11, 30, 34]. These methods aim at leveraging massive noisy labeled data with a small set of clean labels to learn a robust image classifier. [10, 30] proposed a teacher-student framework with a label cleaning network to achieve noisy label learning. Xiao et al. [34] built a probability model to describe the generating process of noisy labels and used the data with clean labels to pre-train the network. Li et al. [15] distilled the knowledge in the clean labels and used it to avoid overfitting to noisy labels. Compared with learning from data with noisy labels solely, clean labels can guide the model to the right direction to some extent. Experimental results in these works showed that even a small set of clean labels have a positive influence on performance improvement. Our method is similar to [15], but we use the data with noisy labels to reduce the risk of overfitting to the data with clean labels under a more generalized framework instead of the single label classification assumption in [15].

Apart from using a small set of clean labels, several studies also introduced side-information (e.g., knowledge graph) to assist in improving model robustness toward noisy labels [3, 15, 33]. Wu et al. [33] used a mixed dependency

Figure 2. An overview of the proposed approach for weakly supervised image classification, which consists of a shared feature extractor (i.e., ResNet-50), a clean net, and a residual net. The clean net and the residual net are responsible for learning a mapping (F_c) from feature space to clean label space and a residual mapping (F_r) from feature space to the residual between clean labels and noisy labels, respectively. We train two classifiers h and g using the noisy labels in the noisy set and the clean labels in the clean set, respectively. h is supervised by the noisy labels y_i for all samples x_i in D_n in terms of L_{noise} , and g is supervised by the clean labels v_j for all samples x_j in D_c in terms of L_{clean} . Residual net r with classifier h works as a regularization term in training the clean net c with classifier g .

graph modeling the semantic hierarchical dependency between labels to complete the missing labels by label propagation. Li et al. [15] introduced a label-to-label graph encoding the structure in label space to avoid over-certainty of the data with clean labels. However, side-information is sometimes highly related to the label set, which also restricts the generalization ability of model to some extent.

3. Our Approach

3.1. Problem Formulation

Our goal is to leverage massive noisy labeled data with a small set of clean labels to obtain a robust image classification model. We also expect that the model does not require assumptions about the label type, and can generalize to both single-label and multi-label image classification tasks.

Let $D = D_n \cup D_c$ denote the entire training dataset, where $D_n = \{(x_i, y_i) | i = 1, \dots, N_n\}$ and $D_c = \{(x_j, v_j) | j = 1, \dots, N_c\}$ are the dataset with noisy labels and the dataset with clean labels, respectively; x_i and y_i in D_n denote the i -th image and corresponding noisy label; N_n denote the total image number in D_n ; x_j and v_j in D_c denote the j -th image and corresponding clean label; N_c denote the total image number of D_c . It should be noted that noisy labels for images in D_c are not required. In this work, we do not make any assumptions about noisy label type, i.e., single-label or multi-label data. In practical applications, we can assume the number of images with clean labels is much less than the noisy data, i.e., $N_c \ll N_n$.

As shown in Fig. 2, we leverage multi-task learning [7, 8, 31] to perform weakly supervised image classification,

which train two classifiers g and h to fit the clean labels in clean set and the noisy labels in noisy set, respectively. The backbone CNN (i.e., a ResNet50 [9] or Inception V3 [27]) is used to learn a mapping from image space x to the feature space f (e.g., pool5). The features are shared by the residual net and the clean net.

Both the clean net and the residual net contain a non-linear transformation which works as an activation layer between two linear layers. The activation layer could be a common used non-linear activate function, such as ReLU, tanh and sigmoid. We will provide the performance with different activation functions in detail in experiments. This non-linear transformation is used to learn a mapping from feature space to clean label space or to noisy label space. The reason why non-linear activation works better than linear is that the shared feature space f may not provide the discriminative ability for both the samples with clean labels and the samples with noisy labels simultaneously.

3.2. Residual Net for Noise Regularization

Classifier g , together with the shared backbone CNN and the clean net, is the final image classifier that the proposed approach aims to learn. It is used to learn a mapping from feature space to clean label space. Let's denote the mapping as F_c and the output of clean net as c ; then c can be represented as

$$c = F_c(f(x)). \quad (1)$$

Similarly, classifier g can be represented as

$$g = (c), \quad (2)$$

where σ is a sigmoid function. Only using the data in clean set to train classifier g makes g is prone to overfit since the size of clean set can be very small in practical scenarios. Therefore, we introduce classifier h , which is expected to serve as a regularization term w.r.t. classifier g .

Specifically, h is used to learn the residual mapping from feature space to the residual between clean labels and noisy labels. Let's denote the residual mapping as F_r and the output of residual net as r ; then r can be represented as

$$r = F_r(f(x)). \quad (3)$$

Similarly, h can be represented as

$$h = \sigma(r + c), \quad (4)$$

where σ is a sigmoid function. In our experiments, we find that summing the values of r and c up before applying sigmoid function helps the network to have a better convergence. So we do the sum operation before sigmoid. We do not need to explicitly distinguish between the multi-label and single label data. Therefore, we use a sigmoid function to generate the probability for both scenarios.

The reason why h can be regarded as a noise regularization term for g is the same as that why use regularization methods, such as weight decay, early stopping and drop out, during network training. They are all helpful to relieve the overfitting problem. Through the above discussion, we can see that the proposed residual net can model the unreliable part in the massive noisy data and then in turn can let classifier g to make use of the reliable part in the massive noisy data to achieve more robust image classification. In this way, the residual net works as a regularization term to relieve the overfitting issue with the classifier g .

The proposed approach learns a mapping from clean label space to noisy label space conditioned on unreliable information in the noisy labels. From the perspective of learning the relation between these two label spaces, it is similar to using label transition model. However, unlike modeling the label transition process explicitly, which usually requires paired noisy-clean label for the samples in clean set, our approach does not require such kind of paired data. Using the clean net and the residual net, our approach can explore the relation between clean labels and noisy labels from unpaired data. Therefore, the two classifiers g and h can be trained separately. This makes it possible for the proposed approach to work under a wide range of applications.

3.3. Network Training

Both h and g are trained with binary cross-entropy loss. The difference lies in the input images are different. h is supervised by the noisy label y_i for all samples i in D_n while g is supervised by the clean label v_j for all samples j in D_c . We denote the loss of h and g as L_{noise} and L_{clean} ,

respectively, and they can be formulated as follows

$$L_{\text{noise}} = -\frac{1}{N_n} \sum_{i \in D_n} (y_i \ln(h_i) + (1 - y_i) \ln(1 - h_i)), \quad (5)$$

$$L_{\text{clean}} = -\frac{1}{N_c} \sum_{j \in D_c} (v_j \ln(g_j) + (1 - v_j) \ln(1 - g_j)), \quad (6)$$

where h_i and g_j are the predictions by classifiers h and g for the corresponding image samples x_i and x_j , respectively.

Given the above definitions, the overall objective during our network training can be formulated as

$$\arg \min_W L_{\text{clean}} + L_{\text{noise}}, \quad (7)$$

where W denotes the parameters of the network and λ denotes the trade-off parameter between two losses. Following [30], to train classifiers g and h jointly leveraging the massive noisy labeled data and a small set of clean labeled data. In each batch during network training, we choose samples from both D_c and D_n in a ratio of 1:9.

We initialize the backbone CNN by the weights of ImageNet pre-trained model. We use different training schemes on multi-label and single-label data. For multi-label image classification, we first fine-tune the backbone CNN by using the noisy labeled data and then only train the clean net and the residual net. For single-label image classification, we directly fine-tune the whole network.

4. Experimental Results

4.1. Datasets

MS COCO2014 dataset [16] is designed for image classification, object detection, and semantic segmentation tasks, which contains about 120K images of 80 classes. We do not use origin MS COCO2014 dataset directly, because it lacks the noisy labels. Following the idea of semi-automatic image annotation in [11, 30], we use an ImageNet [4] pre-trained Inception V3 [27] model to generate annotations for all the images. Specifically, we first map the classes in ImageNet to the classes in MS COCO and remove the classes which do not appear in ImageNet, obtaining a label set with 56 classes, the mapped labels are clean labels. Then we use Inception V3 to generate the top-8 predictions for each image in the origin MS COCO dataset and map them to the 56 label classes. These automatically generated labels can be viewed as the noisy labels. We remove the unlabeled images and finally obtain three sets for train, validation and test, with size of 68,213, 16,714 and 16,763 images, respectively.¹ Among the 68,213 training images, we assume only a small set of images (e.g., 5%) have clean labels during image classification model learning, and all the remaining images only have noisy labels.

¹We plan to put the MS COCO dataset we compiled into public domain.

| Dataset | # Classes | # Train Imgs. | # Val/Test Imgs. |
|-------------|-----------|---------------|------------------|
| MS COCO2014 | 56 | 68K/68K | 16K/16K |
| OpenImage | 6,012 | 9M/40K | -/120K |
| Clothing1M | 14 | 1M/50K | 14K/10K |

Table 1. Evaluations protocols used for the MS COCO2014, OpenImage, and Clothing1M databases.

OpenImage dataset [13] is a public multi-label dataset for image classification. It contains more than 9M images with machine annotated labels from 6,012 unique classes. There are many versions of this dataset since it was proposed. We adopt the first version to evaluate our approach. It contains 9,011,219 images in training set (data with noisy labels) and 167,056 images in validation set (data with both noisy and clean labels). Following the partition in [30], we use the whole training set and a quarter of validation set (about 40K images) to train the model and the remain images in validation set to test.

Clothing1M dataset [34] is a widely used dataset for single-label noise learning. It was proposed by [34] in 2015. It contains 100M clothe images with noisy labels from 14 classes. Different from the noisy label type in OpenImage and the compiling of MS-COCO, which was annotated by pre-train model, labels in Clothing1M are corrupted by real-world noise. According to [34], the noisy label of each image in this dataset is assigned by the keyword of its surrounding text. For the images with clean labels, it was split into train, validation, and test, with size of 50K, 14K and 10K, respectively. We adopt this protocol for consistency with state-of-the-art methods. [14, 23, 34].

4.2. Training Details

The proposed approach and all the baseline approaches are implemented with Tensorflow. We use Inception V3 model as the backbone network for OpenImage and ResNet50 model as the backbone network for MS COCO and Clothing1M. Since MS COCO is a dataset compiled by ourselves, we report the results by the reimplemented baseline methods while using the reported results [30] and [14, 23] on OpenImage and Clothing1M, respectively. On the MS COCO dataset, we first use all the images with noisy/clean labels to train the ResNet50 (Noisy/GT) for 20,000 iterations using a batch size of 64 optimized by RMSProp [29]. The learning rate is initialized with 10^{-4} and decayed by 0.9 every 2 epochs. To evaluate the model performance under a small set of data with clean labels, we use 5%, 10%, and 20% samples from the entire training set to form the clean set. We find the state-of-the-art method by Veit et al. [30] (WP / TJ) is hard to converge on MS COCO when it is optimized by RMSProp. For a fair comparison, we optimize all the models by Adam [12] using a batchsize of 32. We use a learning rate of 10^{-4} to fine-tune the models

for 50,000 addition iterations. We use two pairs of learning rate ($(10^{-4}, 10^{-5})$ for 5%, 10%, $(10^{-3}, 10^{-4})$ for 20%) to train the method by Veit et al. [30] (WP / TJ) for 200,000 iterations. For the proposed method, we use a learning rate of 10^{-4} for sigmoid function and a learning rate of 10^{-5} for tanh and ReLU, and with up to 100,000 iterations. The trade-off parameter empirically set to 0.2.

On the OpenImage dataset, we train the proposed method with a learning rate of 10^{-3} using RMSProp optimizer using a batch size of 32 and up to 2M iterations. We set the trade-off parameter = 0.1. Since not all of the classes in the images of clean set have the clean labels, we just fit the classes which have clean labels. On the Clothing1M dataset, we use a learning rate of 10^{-4} and optimize the model by Adam using a batch size of 32, and up to 120,000 iterations. We set the trade-off parameter = 0.2.

4.3. Metrics

For multi-label image classification, we use the mean of class-average precision (mAP) and the class agnostic average precision (AP_{all}) to evaluate the performance for consistency with [30]. For each binary classification problem, we compute an AP to reflect the prediction precision of positive labels, which can be written as

$$AP_c = \frac{1}{m} \sum_{i=1}^n Pre(i, c) \cdot I(i, c), \quad (8)$$

where m and n denote the number of positive labels and test samples, respectively. $Pre(i, c)$ denotes the precision of class c at rank i . $I(i, c)$ is an indicator function with 1 denoting the positive label's presence of class c at rank i . mAP is the mean of AP_c across all classes while AP_{all} is the AP which treats all classes as a single class by ignoring the class annotations.

For single-label image classification, we follow the state-of-the-art single-label noise learning methods [14, 23], and report the top-1 accuracy.

4.4. Results for Multi-label Image Classification

We first conduct experiments on MS COCO [16] and OpenImage [13] for multi-label image classification. We compare it with the state-of-the-art method [30]. Since the source code of [30] is not publicly available, we implemented it based on our best understanding, and achieves similar results (62.17% mAP and 89.15% AP_{all}) to that in [30] on OpenImage [13]. We use WP and TJ to denote the two variants (with pre-training, trained jointly) in [30], respectively. We also provide the results of several related baselines.

Backbone (Noisy): A backbone network is trained for multi-label classification using all the noisy labels in dataset. This can be viewed as a lower bound of all methods which use clean labels.

| Method \ Dataset | MS COCO | | | | | | OpenImage | |
|-----------------------|--------------|--------------|--------------|-------------------|--------------|--------------|--------------|-------------------|
| | mAP | | | AP _{all} | | | mAP | AP _{all} |
| | 5% | 10% | 20% | 5% | 10% | 20% | | |
| Backbone (Noisy) | 45.30 | | | 39.14 | | | 61.82 | 83.82 |
| Backbone (Noisy-FT-W) | 54.94 | 56.34 | 59.03 | 56.86 | 58.36 | 61.26 | 61.53 | 85.88 |
| Backbone (Noisy-FT-L) | 55.10 | 56.77 | 58.72 | 57.53 | 59.33 | 61.18 | 65.66 | 89.57 |
| Backbone (Noisy-FT-M) | 49.78 | 50.13 | 50.29 | 45.99 | 46.03 | 46.16 | 61.90 | 84.80 |
| Veit et al. [30] (WP) | 46.57 | 49.83 | 51.83 | 47.63 | 50.63 | 52.64 | 62.36 | 87.68 |
| Veit et al. [30] (TJ) | 46.42 | 50.15 | 52.14 | 46.47 | 50.59 | 52.76 | 62.38 | 87.67 |
| Proposed (sigmoid) | 58.28 | 59.51 | 60.55 | 59.86 | 61.12 | 62.34 | 69.02 | 94.08 |
| Proposed (tanh) | 57.95 | 59.19 | 60.11 | 60.03 | 61.39 | 62.33 | – | – |
| Proposed (ReLU) | 58.16 | 59.29 | 60.19 | 60.19 | 61.54 | 62.44 | – | – |
| Backbone (GT) | 65.16 | | | 68.15 | | | – | – |

Table 2. Multi-label classification performance (in %) by the proposed approach and several baseline methods on the MS COCO and OpenImage datasets. We choose 5%, 10% and 20% clean labels in MS COCO and all clean labels in training set of OpenImage as the clean set, respectively. The best results except for the results using 100% clean labels, denoted as Backbone (GT), are highlighted in bold.

Backbone (GT): A backbone network is trained for multi-label classification using all the clean labels in dataset. This can be viewed as an upper bound of all methods which use clean labels. It should be noted that the Backbone (GT) is only trained on MS COCO since the clean labels of whole OpenImage dataset are missing.

Backbone (Noisy-FT-W): Fine-tune the whole network of Backbone (Noisy) with clean labels in the clean set. This method directly uses clean labels to train a large network which is prone to overfit when clean labels are few.

Backbone (Noisy-FT-M): Fine-tune the last layer of Backbone (Noisy) with mixed labels in the clean set. The mixed labels consist of both clean labels in clean set and noisy labels in noisy set (in a ratio of 1 : 9).

Backbone (Noisy-FT-L): Fine-tune the last layer of Backbone (Noisy) with clean labels in the clean set. This method relieves the problem of overfitting by reducing parameters in training.

The results on MS COCO and OpenImage are reported in Table 4.4 in terms of mAP and AP_{all}. We can see that on MS COCO, all of the weakly supervised methods can significantly improve the performance compared to the baseline method - Backbone (Noisy), even using only 5% clean labels. This shows the positive influence of clean labels on noisy label learning, and this influence grows with the increase of clean label quantity. However, when extremely few clean labels are available (e.g., 5%), the improvement of model performance becomes more difficult. Compared to other methods, the proposed method shows the least performance decrease (2.3% by our method vs. 3.6% by Backbone (Noisy-FT-L) in terms of mAP) from using 20% to 5% clean labels while keeps the best performance.

Similar to MS COCO, we also conducted experiments over different percent of clean labels for the training set on OpenImage. For OpenImage dataset, the size of whole clean set in training is about 40K. We then used several sub-

| Metric | Percent of clean labels | | | | | |
|-------------------|-------------------------|-------|-------|-------|-------|-------|
| | 10% | 20% | 40% | 60% | 80% | 100% |
| mAP | 65.08 | 65.98 | 67.20 | 67.97 | 68.61 | 69.02 |
| AP _{all} | 91.08 | 92.18 | 93.13 | 93.65 | 93.86 | 94.08 |

Table 3. Performance of image classification on OpenImage (in %) by the proposed approach using different percent of clean labels for the training set.

sets of the clean set to train the model, with ratio of 10%, 20%, 40%, 60%, and 80%, respectively. The results are given in Table 3. We can see that the proposed method can keep the best classification performance even the clean set size is reduced to 20%. It concurs with the results on MS COCO, which achieves similar performance with Backbone (Noisy-FT-W) by using only half data (59.51% by our method using 10% clean labels vs. 59.03% by Backbone (Noisy-FT-W) using 20% clean labels in terms of mAP). This can be encouraging and has practical significance in real scenarios since image annotation is usually an expensive and time-consuming work, which is more practicable to annotate 8K images rather than 40K images.

The results on MS COCO and OpenImage datasets demonstrate the effectiveness of our model in leveraging massive noisy labeled data with a small set of clean labels to perform weakly supervised learning. The proposed approach achieves 3.1% (on MS COCO by using 5% clean labels) and 3.3% (on OpenImage) higher performance than the best of the baseline methods in terms of mAP. We give some examples of the classification results in Fig. 3. The proposed method performs well in many hard cases that the baseline methods cannot work well. It should be noted that we only train the clean net and the residual net by fixing the backbone network, and it is possible to have better results if the entire network is trained. The possible reasons why our method performs better than the state-of-the-art method

Figure 3. Examples of multi-label image classification by the proposed method and the baseline methods on the test set of the Open-Image and MS COCO datasets. We provide the top-5 most confident predictions by Backbone (Noisy), Backbone (Noisy-FT-L) and the proposed method, denoted as Baseline1, Baseline2 and Proposed, respectively. We choose 5% clean labels as the clean set of MS COCO.

Figure 4. The mAP of image classification on MS COCO by the proposed approach using different activate functions (sigmoid, tanh and ReLU) and 5% clean labels.

by Veit et al. [30] are: i) we use a non-linear transformation to learn the mapping from image feature space to label space while a linear transformation was used in [30]; ii) our approach can make use of all the noisy data of the whole dataset during network training, while [30] only used the noisy labels of the images in the clean set.

We also report the performance using different activate functions, i.e., sigmoid, tanh and ReLU, in order to get comprehensive understanding of the proposed method. The results on MS COCO dataset using three different activation functions are given in Table 4.4. We can see that the performance difference between three activate functions is minor, which is less than 0.5% from 5% to 20% clean labels

| # | Method | Data | Pretrain | top - 1 |
|---|----------------------------------|----------------------|----------|--------------|
| 1 | ResNet50 (Noisy) | Noisy set | ImageNet | 68.94 |
| 2 | ResNet50 (Clean) | Clean set | ImageNet | 75.19 |
| 3 | CleanNet- w_{soft} [14] | Noisy set | ImageNet | 74.69 |
| 4 | Fine-tune | Clean set | #3 | 79.90 |
| 5 | Forward [23] | Noisy set | ImageNet | 69.84 |
| 6 | Fine-tune | Clean set | #5 | 80.38 |
| 7 | Proposed (sigmoid) | Noisy set, Clean set | ImageNet | 79.93 |

Table 4. Single-label classification performance (in %) by the proposed approach and several state-of-the-art methods on the Clothing1M dataset. The best results are highlighted in bold.

in terms of both mAP and AP_{all} . However, we find that the convergence speeds of the three functions are different. We also report the changes of mAP over iterations on MS COCO dataset by using 5% clean labels. As shown in Fig. 4, the sigmoid function has the slowest convergence speed among all three activate functions under the same learning rate (10^{-5}). This is the reason why we use a higher learning rate for the sigmoid function in our experiment.

4.5. Results for Single-label Image Classification

We perform single-label image classification on Clothing1M to evaluate the performance of the proposed approach and provide comparisons with the state-of-the-art methods. Two important baselines we used are CleanNet- w_{soft} [14] and Forward [23]. These methods are designed for learning from the data with only noisy labels. Thus, to utilize the data with clean labels, they need to fine-tune the model, which denoted by Fine-tune in Table 4. Besides the state-of-the-art methods, we also provide two other important traditional baselines, ResNet50 (Noisy) and ResNet50 (Clean), which trains a backbone network using all the data with noisy labels or clean labels in the dataset.

The results are reported in Table 4 in terms of top-1 accuracy. We can see that the proposed method has comparable accuracy with other state-of-the-art methods (79.93% by our method vs. 80.38% by Forward [23] and 79.90% by CleanNet- w_{soft} [14]) on the Clothing1M dataset. Both Forward [23] and the proposed method learn a mapping from clean label space to noisy label space. Compared with the proposed method, Forward [23] introduced extra information, which used the paired noisy-clean labels to estimate the label confusion matrix. This may be the reason why the performance of [23] is slightly higher than us. However, in real applications, paired labels are not available sometimes, e.g., only 25K in 50K images in the training set of Clothing1M dataset have both clean and noisy labels. Thus the requirement on paired labels in Forward [23] may also limit its usage. For the state-of-the-art method [14], although we achieve very similar results to it, the proposed method does not need to generate the reference set which sometimes can be very time-consuming.

Figure 5. Examples of single-label image classification generated by the proposed method on Clothing1M. We show the top-3 predictions by the classifier g and compare it with the corresponding predictions of the classifier h. The green and red labels denote the clean and the noisy label, respectively. We normalize the predictions by both classifiers for easy comparisons.

Compared to the state-of-the-art methods [23, 14], which are restricted to single-label image classification, the merit of the proposed method is that it can be applied to multi-label image classification. The proposed method does not have the single-label noise assumption, and performs well on the multi-label dataset as the experiments showed previously. It suggests that our method generalizes well in both single-label and multi-label image classification scenarios. Another merit of the proposed method is that it can use a 1-step training scheme while Forward [23] and CleanNet- w_{soft} [14] need to first train the base model on the data with noisy labels and then do the fine-tuning on data with clean labels. The proposed method uses the entire data to train the whole network for only one time, which simplifies the training process while achieves high accuracy. Fig. 5 shows the examples of single-label image classification by the proposed method on Clothing1M. From the residual between the prediction of g and h, we can see that the residual net is helpful for modeling the unreliable part in the noisy data, which makes g and h better fit the clean labels and noisy labels, respectively. However, if the images are too difficult to classify, they may also lead to wrong predictions.

4.6. Influence of Residual Net

To show the influence of the residual net in the whole network, we provide ablation experiments to analyze the performance of classifiers g and h. We provide two variants of our model in addition to proposed method (training g, h jointly): i) training h alone (i.e., $\lambda = 0$); ii) training g alone (i.e., only using clean labels to train the model). As shown in Table 5, training h alone does not perform better

| Method \ Dataset | OpenImage | | Clothing1M |
|-----------------------|--------------|-------------------|--------------|
| | mAP | AP _{all} | top - 1 |
| Backbone (Noisy) | 61.82 | 83.82 | 68.94 |
| training h alone | 64.06 | 81.64 | 67.92 |
| training g alone | 67.34 | 93.73 | 75.13 |
| training g, h jointly | 69.02 | 94.08 | 79.93 |

Table 5. Influence of different training strategies to the image classification accuracy (in %). We use sigmoid as the activate function for both OpenImage and Clothing1M datasets.

than Backbone (Noisy) since there is no explicit difference between the clean net and the residual net. When we introduce the data with clean labels to train classifier g, the different roles of the two nets can be recognized. Compared to training g alone, training g, h jointly achieves 1.7% improvement on OpenImage in terms of mAP, and 4.8% improvement on Clothing1M in terms of top-1 accuracy. The results suggest that the reliable information identified by the residual net from massive noisy labeled data can improve the performance of classifier g in both single-label and multi-label image classification tasks.

5. Conclusion

While weakly supervised image classification utilizing massive noisy labeled data is valuable for practical applications where a large clean dataset is not available, it is challenging because of the difficulty in exploiting the underline semantic information from noisy data. We address these issues by proposing a novel end-to-end trainable approach for weakly supervised learning that does not require assumptions about the label type. The proposed approach consists of a clean net and a residual net, which work in a multi-task way, and are responsible for learning a mapping from feature space to clean label space and a mapping from feature space to the residual between clean labels and noisy labels, respectively. As a result, the residual net can serve as a regularization term to reduce the risk of overfitting of the clean net. Extensive evaluations using both multi-label and single-label image classification tasks on the MS COCO, OpenImage, and Clothing1M datasets show that the proposed approach achieves promising results compared with state-of-the-art, and has good generalization ability.

Acknowledgment

This research was supported in part by the National Key R&D Program of China (grant 2017YFA0700800), Natural Science Foundation of China (grants 61672496, 61650202, and 61772500), and External Cooperation Program of Chinese Academy of Sciences (CAS) (grant GJHZ1843).

References

- [1] E. Beigman and B. B. Klebanov. Learning with annotation noise. In *Proc. ACL*, pages 280–287, 2009.
- [2] C. E. Brodley and M. A. Friedl. Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, 11(1):131–167, 2011.
- [3] X. Chen and A. Gupta. Webly supervised learning of convolutional networks. In *Proc. IEEE ICCV*, pages 1431–1439, 2015.
- [4] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE CVPR*, pages 248–255, 2009.
- [5] B. Frénay and M. Verleysen. Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks & Learning Systems*, 25(5):845–869, 2014.
- [6] A. Ghosh, H. Kumar, and P. S. Sastry. Robust loss functions under label noise for deep neural networks. In *Proc. AAAI*, 2017.
- [7] H. Han, A. K. Jain, F. Wang, S. Shan, and X. Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 40(11):2597–2609, Nov. 2018.
- [8] H. Han, C. Otto, X. Liu, and A. K. Jain. Demographic estimation from face images: Human vs. machine performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 37(6):1148–1161, Jun. 2015.
- [9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. IEEE CVPR*, pages 770–778, 2016.
- [10] M. Hu, H. Han, S. Shan, and X. Chen. Multi-label learning from noisy labels with non-linear feature transformation. In *Proc. ACCV*, 2018.
- [11] N. Inoue, E. Simoserra, T. Yamasaki, and H. Ishikawa. Multi-label fashion image classification with minimal human supervision. In *Proc. IEEE ICCV Workshops*, pages 2261–2267, 2017.
- [12] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.
- [13] I. Krasin, T. Duerig, N. Alldrin, A. Veit, S. Abu-El-Haija, S. Belongie, D. Cai, Z. Feng, V. Ferrari, V. Gomes, A. Gupta, D. Narayanan, C. Sun, G. Chechik, and K. Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *arXiv:1701.01619*, 2016.
- [14] K. H. Lee, X. He, L. Zhang, and L. Yang. CleanNet: Transfer learning for scalable image classifier training with label noise. In *Proc. IEEE CVPR*, 2018.
- [15] Y. Li, J. Yang, Y. Song, L. Cao, J. Luo, and L.J. Li. Learning from noisy labels with distillation. In *Proc. IEEE ICCV*, pages 1928–1936, 2017.
- [16] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. In *Proc. ECCV*, pages 740–755, 2014.
- [17] T. Liu and D. Tao. Classification with noisy labels by importance reweighting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38:447–461, 2016.
- [18] W. Liu, G. Hua, and J. R. Smith. Unsupervised one-class learning for automatic outlier removal. In *Proc. IEEE CVPR*, pages 3826–3833, 2014.
- [19] N. Manwani and P. S. Sastry. Noise tolerance under risk minimization. *IEEE Transactions on Cybernetics*, 43(3):1146–1151, 2013.
- [20] I. Misra, C. L. Zitnick, M. Mitchell, and R. Girshick. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *Proc. IEEE CVPR*, pages 2930–2939, 2016.
- [21] N. Natarajan, I. S. Dhillon, P. Ravikumar, and A. Tewari. Learning with noisy labels. In *Proc. NIPS*, pages 1196–1204, 2013.
- [22] L. Niu, Q. Tang, A. Veeraraghavan, and A. Sabharwal. Learning from noisy web data with category-level supervision. In *Proc. IEEE CVPR*, 2018.
- [23] G. Patrini, A. Rozza, A. K. Menon, R. Nock, and L. Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proc. IEEE CVPR*, pages 2233–2241, July. 2017.
- [24] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv:1412.6596*, 2014.
- [25] D. Rolnick, A. Veit, S. Belongie, and N. Shavit. Deep learning is robust to massive label noise. *arXiv:1705.10694*, 2017.
- [26] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus. Training convolutional networks with noisy labels. In *Proc. ICLR Workshops*, 2015.
- [27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proc. IEEE CVPR*, pages 2818–2826, 2016.
- [28] D. Tanaka, D. Ikami, T. Yamasaki, and K. Aizawa. Joint optimization framework for learning with noisy labels. In *Proc. IEEE CVPR*, 2018.
- [29] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4(2), 2012.
- [30] A. Veit, N. Alldrin, G. Chechik, I. Krasin, A. Gupta, and S. Belongie. Learning from noisy large-scale datasets with minimal supervision. In *Proc. IEEE CVPR*, pages 6575–6583, 2017.
- [31] F. Wang, H. Han, S. Shan, and X. Chen. Deep multi-task learning for joint prediction of heterogeneous face attributes. In *Proc. 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 173–179, 2017.
- [32] Y. Wang, W. Liu, X. Ma, J. Bailey, H. Zha, L. Song, and S. T. Xia. Iterative learning with open-set noisy labels. In *Proc. IEEE CVPR*, 2018.
- [33] B. Wu, F. Jia, W. Liu, B. Ghanem, and S. Lyu. Multi-label learning with missing labels using mixed dependency graphs. *International Journal of Computer Vision*, 126(8):875–896, 2018.
- [34] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang. Learning from massive noisy labeled data for image classification. In *Proc. IEEE CVPR*, pages 2691–2699, 2015.