CleanNet: Transfer Learning for Scalable Image Classifier Training with Label Noise

Kuang-Huei Lee\textsuperscript{1} Xiaodong He\textsuperscript{2} Lei Zhang\textsuperscript{1} Linjun Yang\textsuperscript{3*}
\textsuperscript{1}Microsoft AI and Research \textsuperscript{2}Microsoft Research \textsuperscript{3}Facebook
\{kualee,xiaohe,leizhang\}@microsoft.com linjuny@fb.com

Abstract

In this paper, we study the problem of learning image classification models with label noise. Existing approaches depending on human supervision are generally not scalable as manually identifying correct or incorrect labels is time-consuming, whereas approaches not relying on human supervision are scalable but less effective. To reduce the amount of human supervision for label noise cleaning, we introduce CleanNet, a joint neural embedding network, which only requires a fraction of the classes being manually verified to provide the knowledge of label noise that can be transferred to other classes. We further integrate CleanNet and conventional convolutional neural network classifier into one framework for image classification learning. We demonstrate the effectiveness of the proposed algorithm on both of the label noise detection task and the image classification on noisy data task on several large-scale datasets. Experimental results show that CleanNet can reduce label noise detection error rate on held-out classes where no human supervision available by 41.5% compared to current weakly supervised methods. It also achieves 47% of the performance gain of verifying all images with only 3.2% images verified on an image classification task.

1. Introduction

One of the key factors that drive recent advances in large-scale image recognition is massive collections of labeled images like ImageNet \cite{5} and COCO \cite{15}. However, it is normally expensive and time-consuming to collect large-scale manually labeled datasets. In practice, for fast development of new image recognition tasks, a widely used surrogate is to automatically collect noisy labeled data from Internet \cite{6,25,11}. Yet many studies have shown that label noise can affect accuracy of the induced classifiers significantly \cite{27,19,22}, making it desirable to develop algorithms for learning in presence of label noise.

Learning with label noise can be categorized by type of supervision: methods that rely on human supervision and methods that do not. For instance, some of the large-scale training data were constructed using classifiers trained on manually verified seed images to remove label noise (e.g. LSUN \cite{37} and Places \cite{38}). Some studies for learning convolutional neural networks (CNNs) with noise also rely on manual labeling to estimate label confusion \cite{35,20}. The methods using human supervision exhibit a disadvantage in scalability as they require labeling effort for every class. For classification tasks with millions of classes \cite{8,4}, it is infeasible to have even one manual annotation per class. In contrast, methods without human supervision (e.g. model predictions-based filtering \cite{7} and unsupervised outliers removal \cite{34,17,24}) are scalable but often less effective and more heuristic. Going with any of the existing approaches, either all the classes or none need to be manually verified. It is difficult to have both scalability and effectiveness.

In this work, we strive to reconcile this gap. We observe that one of the key ideas for learning from noisy data is finding “class prototypes” to effectively represent classes. Methods learn from manually verified seed images like \cite{17} and methods assume majority correctness like \cite{11} belong to this category. Inspired by this observation, we develop an attention mechanism that learns how to select representative seed images in a reference image set collected for each class with supervised information, and transfer the learned knowledge to other classes without explicit human supervision through transfer learning. This effectively addresses the scalability problem of the methods that rely on human supervision.

Thus, we introduce “label cleaning network” (CleanNet), a novel neural architecture designed for this setting. First, we develop a reference set encoder with the attention mechanism to encode a set of reference images of a class to an embedding vector that represents that class. Second, in parallel to reference set embedding, we also build a query embedding vector for each individual image and impose a matching constraint in training to require a query embedding to be similar to its class embedding if the query is relevant to its class. In other words, the model can tell...
whether an image is mislabeled by comparing its query embedding with its class embedding. Since class embeddings generated from different reference sets represent different classes where we wish the model to adapt to, CleanNet can generalize to classes without explicit human supervision. Fig. 1 illustrates the end-to-end differentiable model.

As the first step of this work, we demonstrate that CleanNet is an effective tool for label noise detection. Simple thresholding based on the similarity between the reference set and the query image lead to good results compared with existing methods. Label noise detection not only is useful for training image classifiers with noisy data, but also has important values in applications like image search result filtering and linking images to knowledge graph entities.

CleanNet predicts the relevance of an image to its noisy class label. Therefore, we propose to use CleanNet to assign weights to image samples according to the image-to-label relevance to guide training of the image classifier. On the other hand, as a better classifier provides more discriminative convolutional image features for learning CleanNet, we will refresh the CleanNet using the newly trained classifier. We introduce a unified learning scheme to train the CleanNet and image classifier jointly.

To summarize, our contributions include a novel neural architecture CleanNet that is designed to make label noise detection and learning from noisy data with human supervision scalable through transfer learning. We also propose an unified scheme for training CleanNet and the image classifier with noisy data. We carried out comprehensive experimentation to evaluate our method for label noise detection and image classification on three large datasets with real-world label noise: Clothing1M [35], WebVision [13], and Food101-N. Food101-N contains 310K images we collected from Internet with the Food101 taxonomy [2], and we added “verification label” that verifies whether a noisy class label is correct for an image. Experimental results show that CleanNet can reduce label noise detection error rate on held-out classes where no human supervision available by 41.5% compared to current weakly supervised methods. It also achieves 47% of the performance gain of verifying all images with only 3.2% images verified on an image classification task.

2. Related Work

Label noise reduction. Our method belongs to the category of approaches that address label noise by demoting or removing mislabeled instances in training data. One of the popular approaches is unsupervised outlier removal (e.g., One-Class SVM [24], UOCL [17], and DRAE [34]). Using this approach for label noise detection relies on an assumption that outliers are mislabeled. However, outliers are often not well defined, and therefore removing them presents a challenge [17]. Another approach that also need no human supervision is weakly supervised label noise reduction [7]. For example, Thongkam et al. [29] proposed a classification filtering method that learns an SVM from noisy data and removes instances misclassified by the SVM. Weakly supervised methods are often heuristic and we are not aware of any large dataset actually built with these methods. On the other hand, label noise reduction using human supervision has been widely studied for dataset constructions. For instance, Yu et al. [57] proposed manually labeling seed images and then training multilayer perceptrons (MLPs) to remove mislabeled images. Similarly, the Places dataset [38] was constructed using an AlexNet [12] trained on manually verified seed images. However, methods using human supervision exhibit a disadvantage in scalability as they require human supervision for every class to be cleansed.

Direct neural network learning with label noise. Apart from removing label noise, some algorithms were developed for directly learning neural network with label noise. Several works proposed learning models in the case of having label noise [27, 35, 20], and [35, 20] rely on manual labeling to estimate label confusion for real-world label noise. Azadi et al. [1] developed a regularization method to actively select image features for training, but it depends on features pre-trained for other tasks and hence is less effective. Zhuang et al. [41] proposed attention in random sample groups but did not compare with standard CNN classi-
fiers, and thus is less practical. On the other hand, from a training technique perspective, Rolnick et al. [22] suggested that tuning learning rate and batch size can alleviate the accuracy drop caused by label noise. Several works [8, 32, 14, 26] proposed using side information like knowledge graph, too, which is beyond the scope of this paper.

**Transfer learning with neural network.** There is a large body on literature of learning neural joint embeddings for transfer learning [8, 26, 20, 23, 33]. Tsai et al. [30] trained visual-semantic embeddings with supervised and unsupervised objectives using labeled and unlabeled data to improve robustness of embeddings for transfer learning. Recently Liu et al. [16] and Tzeng et al. [31] exploited adversarial objectives for domain adaptation. Inspired by [30], we also incorporate unsupervised objectives in this work.

### 3. Scalable Learning with Label Noise

We focus on learning an image classifier from a set of images with label noise using transfer learning. Specifically, assume we have a dataset of \( n \) images, i.e., \( X = \{(x_1, y_1), \ldots, (x_n, y_n)\} \), where \( x_i \) is the \( i \)-th image and \( y_i \in \{1, \ldots, L\} \) is its class label, where \( L \) is the total number of classes. Note that the labels are noisy, means some of images’ labels are incorrect.

In this section, we will present the CleanNet, a joint neural embedding network, which only requires a fraction of the classes being manually verified to provide the knowledge of label noise that can be transferred to other classes. We then integrate CleanNet and conventional convolutional neural network (CNN) into one system for image classifier training with label noise. Specifically, we introduce the designs and properties of CleanNet in Section 3.1. In Section 3.3, we integrate CleanNet and the CNN into one framework for image classifier learning from noisy data.

#### 3.1. CleanNet

The overall architecture of CleanNet is shown in Fig. 1. It consists of two parts: a reference set encoder and a query encoder. The reference set encoder \( f_s(\cdot) \) learn to focus on representative features from a noisy reference image set, which is collected for a specific class, and output a class-level embedding vector. Since using all the images in the reference set is computationally expensive, we first create a representative subset, and extract one visual feature vector from each image in that subset to form a representative feature vector set, i.e., \( V^*_c \) denotes the representative reference feature vector set for class \( c \) (reference feature set)

We explored two pragmatic approaches to select \( V^*_c \). The first one is random sampling a subset from all images in class \( c \) and extract features using a pre-trained CNN \( f_u(\cdot) \) as shown in Fig. 1. The second approach is running K-means on the extracted features of all images in class \( c \) to find \( K \) cluster centroids and use them as \( V^*_c \). The K-means step is ignored in the figures. Since the K-means approach shows slightly better result on a held-out set, we choose it for experiments hereafter. We select 50 feature vectors to form \( V^*_c \).

In parallel to reference set encoder, we also develop a query encoder \( f_q(\cdot) \). Let \( q \) denote a query image labeled as class \( c \). The query encoder \( f_q(\cdot) \) maps the query image feature \( v^q = f_u(q) \) to a query embedding \( \phi_v = f_q(v^q) \). We impose a matching constraint such that the query embedding \( \phi_v \) is similar to its class embedding \( \phi_v^c = f_s(V^*_c) \) if the query \( q \) is relevant to its class \( c \). In other words, we decide whether a query is mislabeled by comparing its query embedding vector with its class embedding vector. Since the class labels are noisy, we can further mark up a query image and its class label by a manual “verification label”. The verification label for each image is defined as

\[
\ell = \begin{cases} 
1 & \text{if the image is relevant to its noisy class label} \\
0 & \text{if the image is mislabeled} \\
-1 & \text{if verification label not available}
\end{cases}
\]

Note that, to reduce human labeling effort, most of the verification labels are -1, means no human verification available.

The model learns the matching constraint from supervised information in the verification labels, such that the query embedding is similar to its class embedding if the query image \( q \) truly belongs to its class label, and transfer to different classes where no human verification available. In the following, we will present how we build the reference set encoder, query encoder, and objectives for learning the matching constraint.

**Reference set encoder.** The architecture of the reference set encoder is depicted in Fig. 2. It maps a reference feature set \( V^*_c \) for class \( c \) to a class embedding vector \( \phi_v^c \). First, a two-layer MLP projects each image feature to a hidden representation \( h_i \). Next, we learn an attention mechanism to encode representative features to a fixed-length hidden
Figure 3. Examples that received the most and the least attention in a reference set for "cup cakes". $\alpha_i$ is defined in Eq. (4).

representation as class prototype:

$$u_i = \tanh(W h_i + b)$$  \hspace{1cm} (2)

$$\alpha_i = \frac{\exp(u_i^T u)}{\sum_i \exp(u_i^T u)}$$  \hspace{1cm} (3)

$$h = \sum \alpha_i h_i$$  \hspace{1cm} (4)

As shown in Eq. (4), the importance of each $h_i$ is measured by the similarity between $u_i$ and a context vector $u$. Similar to [36], the context vector $u$ is learned during training. Driven by the matching constraint, this attention mechanism learns how to pay attention on the most representative features for classes. This model will learn from supervised information, i.e., the manual verification label, and adapt to other classes without explicit supervision. An example of this attention mechanism is shown in Fig. 3. Finally, a one-layer MLP maps the hidden representation to the class embedding $\phi_c$.

**Query encoder.** As illustrated in Fig. 1, we adopt a 5-layer autoencoder [10] as the query encoder and incorporate autoencoder reconstruction error into learning objectives. Taking this strategy, as proposed in [30], forces the query embedding to preserve semantic information of all the classes including those classes without verification labels, because images without verification label can now be used in training with this unsupervised objective. It has been proven effective for improving domain adaptation performance.

Given a query image feature vector $v^q$, the autoencoder maps $v^q$ to a hidden representation $\phi^g$ and seek to reconstruct $v^q$ from $\phi^g$. The reconstruction error is defined as

$$L_r(v^g) = ||v^q - r(v^g)||^2$$  \hspace{1cm} (5)

where $r(v^g)$ is the reconstructed representation.

**Learning objectives based on matching constraint.** With the supervised information from human verification labels, the similarity between class embedding $\phi^c$ and query embedding $\phi^g$ is maximized if a query is relevant to its class ($l = 1$); otherwise the similarity is minimized ($l = 0$). We adopt the cosine similarity loss with margin to impose this constraint:

$$L_{cos}(\phi^g, \phi^c, l) = \begin{cases} 1 - \cos(\phi^g, \phi^c) & \text{if } l = 1 \\ \omega(\max(0, \cos(\phi^g, \phi^c) - \rho)) & \text{if } l = 0 \\ 0 & \text{if } l = -1 \end{cases}$$  \hspace{1cm} (6)

where $\cos(\cdot)$ is the normalized cosine similarity, $\omega$ is negative sample weight for balancing positive and negative samples, and $\rho$ is the margin set to 0.1 in this work. The case $l = -1$ is ignored in the loss function since this supervised objective only utilizes query images with verification label.

On the other hand, images without verification label can also be utilized to learn the matching constraint. Similar to [30], we introduce an unsupervised self-reinforcing strategy that applies pseudo-verification to images without verification label. To be specific, a query is treated as relevant if $\cos(\phi^g, \phi^c)$ is larger than the margin $\rho$:

$$L_{unsup}^{cos}(\phi^g, \phi^c, l) = \begin{cases} 1 - \cos(\phi^g, \phi^c) & \text{if } l_{sudo} = 1 \\ 0 & \text{if } l_{sudo} = 0 \end{cases}$$  \hspace{1cm} (7)

$$l_{sudo} = \begin{cases} 1 & \text{if } \cos(\phi^g, \phi^c) \geq \rho \\ 0 & \text{otherwise} \end{cases}$$  \hspace{1cm} (8)

where $\rho$ is the same margin as in Eq. (6). From Eq. (7) and Eq. (8), we can see that for queries that are initially treated as relevant, the model will learn to further push up the similarity between queries and reference sets; for queries that are initially treated as irrelevant, they will be ignored.

**Total loss.** To summarize the training objectives, our model is learned by minimizing a total loss combining both supervised and unsupervised objectives:

$$L_{total} = L_{cos} + \beta L_r + t \gamma L_{unsup}^{cos}$$  \hspace{1cm} (9)

$$t = \begin{cases} 1 & \text{if } l = -1 \\ 0 & \text{if } l \in \{0, 1\} \end{cases}$$  \hspace{1cm} (10)

where $\beta$ and $\gamma$ are selected through hyper-parameter search, and $t$ indicates whether a query image has verification label. $\beta$ and $\gamma$ are set to 0.1 in this work. During training, we randomly sample images without verification label as queries for a fraction of a mini-batch (usually 1/2).

Note that the parameters of the attentional reference set encoder and the query encoder are tied across all classes so the information learned from classes that have human verification labels can be transferred to other classes that have no human verification label.

**3.2. CleanNet for Label Noise Detection.**

From a relevance perspective, CleanNet can be used to rank all the images with label noise for a class by cosine similarity $\cos(\phi^g, \phi^c)$. We can simply perform thresholding for label noise detection:

$$\hat{l} = \begin{cases} 1 & \text{if } \cos(\phi^g, \phi^c) \geq \delta \\ 0 & \text{otherwise} \end{cases}$$  \hspace{1cm} (11)
where $\delta$ is a threshold selected through cross-validation. We observe that the threshold is not very sensitive to different classes in most cases, and therefore we usually select an uniform threshold for all classes so that verification labels are not required for all classes for cross-validation.

### 3.3. CleanNet for Learning Classifiers

CleanNet predicts the relevance of an image to its noisy class label by comparing the query embedding of the image to its class embedding that represents the class. That is, the distance between two embeddings can be used to decide how much attention we should pay to a data sample in training the image classifier. Specifically, we assign attention weights on data samples based on the cosine similarity:

$$w_{soft}(x, y = c, V^c_v) = \max(0, \cos(f_q(x)), f_s(V^c_v)))$$  \hspace{1cm} (12)

where $V^c_v$ is the reference image feature set that represents the prototype of class $y = c$. Eq. (12) defines a soft weighting on an image $x$ with noisy class label $y = c$. Similarly, we also define a hard weighting as

$$w_{hard}(x, y = c, V^c_v) = \begin{cases} 1 & \text{if } \cos(f_q(x)), f_s(V^c_v)) \geq \delta \\ 0 & \text{otherwise} \end{cases}$$  \hspace{1cm} (13)

where $\delta$ is a threshold as in Eq. (11). In essence, hard weighting is equivalent to explicit label noise removal. With $w_{soft}$ or $w_{hard}$, we define the weighted classification learning objective as

$$L_{weighted}(x, y = c, V^c_v) = w_{soft/hard}(x, y, V^c_v)H(x, y = c)$$  \hspace{1cm} (14)

where $H(x, y = c)$ is negative log likelihood:

$$H(x, y = c) = -\sum_{c=0}^{L} p(y = c|x)\log p(y = c|x)$$  \hspace{1cm} (15)

### Integrating CleanNet and the image classifier

Learning the image classifier relies on CleanNet to assign proper attention weights to data samples. On the other hand, better classifier provides more discriminative features which are critical for CleanNet learning. Therefore, we integrate CleanNet and the CNN-based image classifier into one framework for end-to-end learning of image classifiers with label noise. The overall architecture of this framework is illustrated in Fig. 4. The structure of a CNN-based image classifier is split into fully-connected layer(s) and convolutional layers $f_c$ that can be used for feature extraction.

### Alternating training

We adopt an alternating training scheme to learn the proposed classification system. At step 1, we first train a classifier from noisy data with all sample weights set to 1. At step 2, parameters of convolutional layers $f_c$ are copied to feature extractor $f_v$ and a CleanNet is trained to convergence. At step 3, the classifier are fine-tuned using the sample weights proposed by CleanNet.

### 4. Experiments

#### 4.1. Datasets

Table I lists the statistics of the datasets.

**Food101-N**: We collect 310k images from Google, Bing, Yelp, and Tripadvisor using the Food101 taxonomy, and avoid foodspotting.com where the original Food101 was collected. The estimated noisy class label accuracy is 80%. We manually add 55k verification label for training and 5k for testing label noise detection. We will release this dataset to community. Image classification is evaluated on Food101 test set.

**Clothing1M** [35]: Clothing1M is the only public large-scale dataset designed for learning from noisy data with human supervision. It consists of 1M images with noisy class labels from 14 fashion classes. The estimated accuracy of
Table 2. Label noise detection in terms of average error rate over all the classes (%). CleanNet* denotes the results using image features extracted from the classifiers retrained with data cleansed by CleanNet.

class labels is 61.54%. There are also three sets of images, with the size of 50k, 14k, 10k, respectively, which have correct class labels provided by human labelers, which we call them clean sets. There are some images overlap between the three clean sets and the noisy set. For those overlapped images, we can then verify whether the noisy class label (as in the noisy set) is correct or not given the human labels on these images, and hence obtained verification labels for these images. Through this process, we obtain 25k and 7k verification labels for training and validation, respectively. The state of the art result of image classification on Clothing1M was reported in [20].

WebVision[13]: WebVision contains 2.4M noisy labeled images crawled from Flickr and Google using the ILSVRC taxonomy [5]. We conveniently verify noisy class labels using the Inception-ResNet-V2 model [28] pre-trained on ILSVRC. Noisy class label of an image is verified as relevant if it falls in top-5 predictions. Otherwise, the noisy class label is verified as mislabeled. We randomly obtain 250 “pseudo-verification labels” for each class for training. For evaluating image classification, we use 50k WebVision validation set and 50k ILSVRC 2012 validation set.

4.2. Label Noise Detection

We first evaluate CleanNet for the task of label noise detection. The label noise detection problem can be viewed as a binary classification problem for each class, and hence the results and comparisons are reported in average error rate over all the classes. We compare with the following categories of existing baseline methods:

- Supervised: Supervised methods learn a binary classification from verification labels for each class. We consider neural networks (2-layer MLP, used in [37] for data construction), kNN, SVM, label prop [40], and label spread [39]. We also explored MLPs of more layers but 2-layer shows the best results.
- Unsupervised: We consider DRAE [34], the state of the art unsupervised outlier removal. Empirically, DRAE shows better results than one-class SVM [24].
- Weakly supervised: Like unsupervised method, weakly supervised methods do not require verification labels. We compare with a widely used classification filtering method: we train a CNN model on noisy data and predict top-K classes for each training image. An image is classified as relevant to its class label if the class is in top-K predictions. Otherwise, it is classified as mislabeled. K is selected on the validation set.

We provide two additional baselines: naive baseline that treats all class labels as correct, and Average baseline that simply averages reference features as a class embedding vector and use query feature as a query embedding vector.

CleanNet and all the baselines depend on a CNN to extract image features. We fine-tune the ImageNet pre-trained ResNet-50 models [9] on noisy data, same as step 1 in the alternating training scheme, and extract the pool5 layer as image features. Implementations of kNN, SVM, label prop, and label spread are from scikit-learn [21]. We re-implemented DRAE and MLP in our experimentation.

In the following, we will evaluate CleanNet for label noise detection under two scenarios: Full supervision: verification labels in all classes are available for learning CleanNet; Transfer learning: only a fraction of classes contain verification labels for learning CleanNet.

Full supervision. In Table 2 we report the label noise detection results in terms of average error rate over all the classes. CleanNet gives error rate of 9.61% on Food101-N and 15.91% on Clothing1M. Comparing to MLP at 10.42% on Food101-N and 16.09% on Clothing1M, we validate that CleanNet performs similar to the best supervised baseline. Comparing to classification filtering at 16.60% on Food101-N and 23.55% on Clothing1M, the results demonstrate effectiveness of adding verification labels for human supervision for label noise detection. CleanNet* denotes the results of CleanNet using image features extracted from the classifiers retrained with data cleansed by CleanNet, and shows improvements (6.99% on Food101-N and 15.77% on Clothing1M). However, improvements become negligible with more iterations.

Transfer learning. We choose Food101-N to demonstrate label noise detection with CleanNet under the setting of transfer learning, where verification labels in n classes are held out for CleanNet. Here we also consider MLP that uses
all verification labels and classification filtering that needs no verification labels. We ONLY evaluate the results on \( n \) held-out classes to demonstrate the results on classes without explicit human supervision. The results are shown in Fig. 5. First, we observe that CleanNet can reduce label noise detection error rate on held-out classes where no human supervision available by 41.5% relatively \((n = 10)\) compared to classification filtering. CleanNet consistently outperforms classification filtering, the weakly-supervised baseline. We also observe that the result of CleanNet with 50/101 classes held out \((11.02\%)\) is still comparable to the result of MLP which is based on supervised learning \((10.12\%)\).

### 4.3. Learning Classifiers with Label Noise

In this subsection, we present experiments for learning image classification models with label noise using the proposed CleanNet-based learning framework. Experimentation in this section is based on ResNet-50.

#### Experiments on Food101-N

Table 3 lists the results on Food101-N using verification labels in all classes. We observe that the performance of smooth soft weighting \((w_{soft})\) \((83.95\%)\) without need for thresholding outperforms hard weighting \((w_{hard})\) \((83.47\%)\). Fig. 6 presents the results of image classification using the proposed CleanNet-based method when verification labels in \( n \) classes are held out. For these \( n \) held-out classes, the information needed for cleaning up the noisy class labels are transferred from other classes through CleanNet. It is observed that there are still 2.1% and 1.75% accuracy gain when 50/101 and 70/101 classes are held out. This validates that labeling effort on a small fraction of classes can still lead to significant gains.

Fig. 6 shows examples of CleanNet. The cosine similarity score between the image and its class’s reference set is shown for each example. Because of transfer learning, CleanNet can assign reasonable scores for images of classes that have not seen in training.

### Table 3. Image classification on Food101-N in terms of top-1 accuracy (%). Verification labels in all classes are available. “None” denotes classifier without any method for label noise.

| method          | data    | top-1 accuracy |
|-----------------|---------|---------------|
| None            | Food101 | 81.67         |
| None            | Food101-N | 81.44        |
| CleanNet, \(w_{hard}\) | Food101-N | 83.47        |
| CleanNet, \(w_{soft}\)    | Food101-N | **83.95**    |

Figure 7. Image classification on Food101-N in terms of top-1 accuracy (%). Red line shows the results when verification labels in \( n \) classes are held out for CleanNet. The blue dashed line shows the baseline without using CleanNet.
Experiments on Clothing1M. For Clothing1M, we consider the state of the art result reported in [20], which also used ResNet-50. [20] used the part of data in Clothing1M that has both noisy and correct class labels to estimate confusion among classes and modeled this information in loss function. Since we only compare the noisy class label to the correct class label for an image to verify whether the noisy class label is correct, we lose the label confusion information, and thus these numbers are not directly comparable. However, labeling the correct classes like Clothing1M (only 14 classes) is not scalable in number of classes because having labeling workers select from a large number of classes is time-consuming and unlikely to be accurate.

Table 5 lists the results of image classification using verification labels in all classes. Using CleanNet significantly improves the accuracy from 68.94% (#1) to 74.69% (#6) on 1M noisy training data. We also follow [20] to fine-tune the best model trained on 1M noisy set on the 50k clean training set. Our proposed method achieves 79.90%, which is comparable to the state of the art 80.30% reported in [20] which benefits from the extra label confusion information.

Experiments on WebVision. As opposed to Food101-N and Clothing1M which are fine-grained tasks, WebVision experiments sheds light on general image classification at very large scale. As mentioned in Sec. 4.1 the pseudo-verification labels are model-based so that we can obtain

| # | method | data | pretrained | top-1 |
|---|--------|------|-----------|-------|
| 1 | None [20] | 1M noisy | ImageNet | 68.94 |
| 2 | None [20] | 50k clean | ImageNet | 75.19 |
| 3 | loss correct. [20] | 1M noisy | ImageNet | 69.84 |
| 4 | None [20] | 50k clean | #3 model | 80.38† |
| 5 | CleanNet, $w_{hard}$ | 1M noisy | ImageNet | 74.15 |
| 6 | CleanNet, $w_{soft}$ | 1M noisy | ImageNet | 74.69 |
| 7 | None | 50k clean | #6 model | 79.90 |

Table 5. Verification conditions: selecting different classes for adding verification labels. Other than every-image, all other conditions have only 250 verification labels in each class.

for all images. This property allows us to explore how to select classes for adding verification labels and compare to the upper bound scenario where all noisy class labels are verified without any cost. We define how to add verification labels as “verification conditions”, listed in Table 5. Table 6 shows the experimental results using CleanNet and soft weighting ($w_{soft}$). We observe that verifying every image (every-image) improves the top-1 accuracy from 58.88% to 63.43% on ILSVRC validation set. With only 3.20% and 1.2% images verified, semantic-308 and random-118 give 47% and 29% of the performance gain of every-image on ILSVRC validation set respectively. Note that we only include 250 verification labels for each class for all experiments using CleanNet. The results again confirm that labeling on a fraction of classes is effective because of transfer learning by CleanNet.

5. Conclusion

In this work, we highlighted the difficulties of having both scalability and effectiveness of human supervision for label noise detection and classification learning from noisy data. We introduced CleanNet as a transfer learning approach to reconcile the issue by transferring supervised information of transferring the correctness of labels to classes without explicit human supervision. We empirically evaluate our proposed methods on both general and fine-grained image classification datasets. The results show that CleanNet outperforms methods using no human supervision by a large margin when small fraction of classes are manually verified. It also matches existing methods that require extensive human supervision when sufficient classes are manually verified. We believe this work creates a novel paradigm that efficiently utilizes human supervision to better address label noise in large-scale image classification tasks.
Acknowledgement

We would like to thank Xi Chen, Bosco Chiu, Yandong Guo and Po-Sen Huang for their thoughtful feedback and discussions. Thanks also to Kelly Huang and Arun Sacheti for helping development of the Food101-N dataset.

References

[1] S. Azadi, J. Feng, S. Jegelka, and T. Darrell. Auxiliary image regularization for deep cnns with noisy labels. arXiv preprint arXiv:1511.07069, 2015.

[2] L. Bossard, M. Guillaumin, and L. Van Gool. Food-101–mining discriminative components with random forests. In European Conference on Computer Vision, pages 446–461. Springer, 2014.

[3] X. Chen and A. Gupta. Webly supervised learning of convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, pages 1431–1439, 2015.

[4] T. Dean, M. A. Ruzun, M. Segal, J. Shlens, S. Vijayanarasimhan, and J. Yagnik. Fast, accurate detection of 100,000 object classes on a single machine. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1814–1821, 2013.

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

[6] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from internet image searches. Proceedings of the IEEE, 98(8):1453–1466, 2010.

[7] B. Frénay and M. Verleysen. Classification in the presence of label noise: a survey. IEEE Transactions on Neural Networks and Learning Systems, 25(5):845–869, 2014.

[8] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In Advances in Neural Information Processing Systems, pages 2121–2129, 2013.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 770–778, 2016.

[10] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. science, 313(5786):504–507, 2006.

[11] J. Krause, B. Sapp, A. Howard, H. Zhou, A. Toshev, T. Duerig, J. Philbin, and L. Fei-Fei. The unreasonable effectiveness of noisy data for fine-grained recognition. In European Conference on Computer Vision, pages 301–320. Springer, 2016.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems, pages 1097–1105, 2012.

[13] W. Li, L. Wang, W. Li, E. Agustsson, and L. Van Gool. Webvision database: Visual learning and understanding from web data. arXiv preprint arXiv:1708.02862, 2017.

[14] Y. Li, J. Yang, Y. Song, L. Cao, J. Luo, and J. Li. Learning from noisy labels with distillation. arXiv preprint arXiv:1703.02391, 2017.

[15] 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 European Conference on Computer Vision, pages 740–755. Springer, 2014.

[16] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In Advances in Neural Information Processing Systems, pages 469–477, 2016.

[17] W. Liu, G. Hua, and J. R. Smith. Unsupervised one-class learning for automatic outlier removal. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3826–3833, 2014.

[18] G. A. Miller. Wordnet: a lexical database for english. Communications of the ACM, 38(11):39–41, 1995.

[19] D. F. Nettleton, A. Orriols-Puig, and A. Fornells. A study of the effect of different types of noise on the precision of supervised learning techniques. Artificial Intelligence Review, 33(4):275–306, 2010.

[20] G. Patrini, A. Rozza, A. Menon, R. Nock, and L. Qu. Making neural networks robust to label noise: a loss correction approach. arXiv preprint arXiv:1609.03683, 2016.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. Journal of Machine Learning Research, 12(Oct):2825–2830, 2011.

[22] D. Rolnick, A. Veit, S. Belongie, and N. Shavit. Deep learning is robust to massive label noise. arXiv preprint arXiv:1705.10694, 2017.

[23] A. Salvador, N. Hynes, Y. Aytar, J. Marin, F. Offi, I. Weber, and A. Torralba. Learning cross-modal embeddings for cooking recipes and food images. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 720:619–508.

[24] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. Neural computation, 13(7):1443–1471, 2001.

[25] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(4):754–766, 2011.

[26] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In Advances in Neural Information Processing Systems, pages 935–943, 2013.

[27] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus. Training convolutional networks with noisy labels. arXiv preprint arXiv:1406.2080, 2014.

[28] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In AAAI, pages 4278–4284, 2017.

[29] J. Thongkam, G. Xu, Y. Zhang, and F. Huang. Support vector machine for outlier detection in breast cancer survivability prediction. Advanced Web and Network Technologies, and Applications, pages 99–109, 2008.
[30] Y.-H. H. Tsai, L.-K. Huang, and R. Salakhutdinov. Learning robust visual-semantic embeddings. arXiv preprint arXiv:1703.05908, 2017.

[31] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. arXiv preprint arXiv:1702.05464, 2017.

[32] A. Veit, N. Alldrin, G. Chechik, I. Krasin, A. Gupta, and S. Belongie. Learning from noisy large-scale datasets with minimal supervision. arXiv preprint arXiv:1701.01619, 2017.

[33] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al. Matching networks for one shot learning. In Advances in Neural Information Processing Systems, pages 3630–3638, 2016.

[34] Y. Xia, X. Cao, F. Wen, G. Hua, and J. Sun. Learning discriminative reconstructions for unsupervised outlier removal. In Proceedings of the IEEE International Conference on Computer Vision, pages 1511–1519, 2015.

[35] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang. Learning from massive noisy labeled data for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2691–2699, 2015.

[36] Z. Yang, D. Yang, C. Dyer, X. He, A. J. Smola, and E. H. Hovy. Hierarchical attention networks for document classification. In HLT-NAACL, pages 1480–1489, 2016.

[37] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365, 2015.

[38] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.

[39] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In Advances in Neural Information Processing Systems, pages 321–328, 2004.

[40] X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107, 2002.

[41] B. Zhuang, L. Liu, Y. Li, C. Shen, and I. Reid. Attend in groups: a weakly-supervised deep learning framework for learning from web data. IEEE Conference on Computer Vision and Pattern Recognition, 2017.