CurriculumNet: Weakly Supervised Learning from Large-Scale Web Images

Sheng Guo, Weilin Huang*, Haozhi Zhang, Chenfan Zhuang, Dengke Dong, Matthew R. Scott, and Dinglong Huang

Malong Technologies, Shenzhen, China
Shenzhen Malong Artificial Intelligence Research Center, Shenzhen, China
{sheng,whuang,haozhang,fan,dongdk,mscott,dlong}@malong.com

Abstract. We present a simple yet efficient approach capable of training deep neural networks on large-scale weakly-supervised web images, which are crawled rawly from the Internet by using text queries, without any human annotation. We develop a principled learning strategy by leveraging curriculum learning, with the goal of handling massive amount of noisy labels and data imbalance effectively. We design a new learning curriculum by measuring the complexity of data using its distribution density in a feature space, and rank the complexity in an unsupervised manner. This allows for an efficient implementation of curriculum learning on large-scale web images, resulting in a high-performance CNN model, where the negative impact of noisy labels is reduced substantially. Importantly, we show by experiments that those images with highly noisy labels can surprisingly improve the generalization capability of model, by serving as a manner of regularization. Our approaches obtain the state-of-the-art performance on four benchmarks, including Webvision, ImageNet, Clothing-1M and Food-101. With an ensemble of multiple models, we achieve a top-5 error rate of 5.2% on the Webvision challenge [18] for 1000-category classification, which is the top performance that surpasses other results by a large margin of about 50% relative error rate. Codes and models are available at: https://github.com/guoshengcv/CurriculumNet.

Keywords: Curriculum learning · weakly supervised · noisy data · large-scale · web images

1 Introduction

Deep convolutional networks have rapidly advanced numerous computer vision tasks, providing the state-of-the-art performance on image classification [9, 31, 34, 14, 37, 8], object detection [28, 27, 22, 20], sematic segmentation [23, 11, 4, 10], etc. They produce strong visual features by training the networks in a fully-supervised manner using large-scale manually annotated datasets, such as ImageNet [5], MS-COCO [21] and PASCAL VOC [6]. Obviously, full and clean human annotations are of crucial importance to achieving a high-performance

* Weilin Huang is the corresponding author (e-mail:whuang@malong.com).

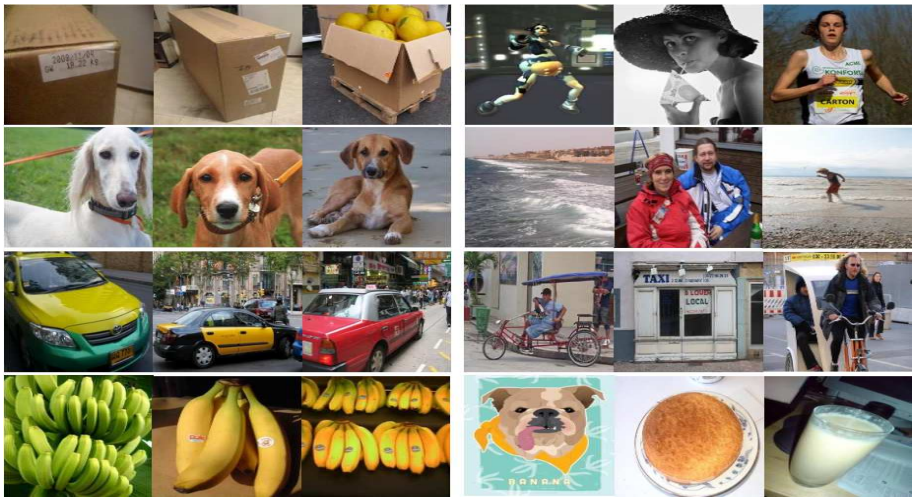

Fig. 1. Image samples of WebVision dataset [19] from the categories of *Carton*, *Dog*, *Taxi* and *Banana*. The dataset was collected from the Internet by using text enquires generated from the 1, 000 semantic concepts of the Imagenet benchmark [5]. Obviously, each category includes a number of mislabeled images as shown on the right.

model, and better results can be reasonably expected if a larger dataset is provided with clean annotations. However, obtaining massive and clean annotations are extremely expensive and time-consuming, making the capability of deep models unscalable to the size of collected data. Furthermore, it is particularly hard to collect clean annotations for tasks where expert knowledge is required, and labels provided by different annotators are possibly inconsistent.

An alternative solution is to use the web as a source of data and supervision, where a large amount of web images can be collected automatically from the Internet by using input queries, such as text information. These queries information can be considered as natural annotations of the images, providing weak supervision of the collected data, which is a cheap way to increase the scale of dataset near-infinitely. However, such annotations are highly unreliable, and often include massive noisy labels. Past work has shown that these noisy labels could significantly affect the performance of deep neural networks on image classification [39]. To address this problem, recent approaches have been developed by proposing robust algorithms against noisy labels [30]. Another solution is to develop noise-cleaning methods that aim to remove or correct the mislabelled examples in training data [32]. However, the noise-cleansing methods often suffer from the main difficulty in distinguishing mislabeled samples from hard samples, which are critical to improving model capability. Besides, semi-supervised methods have also been introduced by using a small subset of manually-labeled images, and then the models trained on this subset are generalized to a larger dataset with unlabelled or weakly-labelled data [36]. Unlike these approaches, we do not aim to propose a noise-cleaning, noise-robust or semi-supervised al-

gorithm. Instead, we investigate improving model capability of standard neural networks by introducing a new training strategy.

In this work, we study the problem of learning convolutional networks from large-scale images with a large amount of noisy labels, such as WebVision challenge [18], which is a 1000-category image classification task having same categories as ImageNet [5]. The labels are provided by simply using the enquires text generated from the 1,000 semantic concepts of the ImageNet [5], *without any manual annotation*. Several image samples are presented in Fig. 1. Our goal is to provide a solution able to handle massive noisy labels and data imbalance effectively. We design a series of experiments to investigate the impact of noisy labels to the performance of deep networks, when the amount of training images is sufficiently large. We develop a simple but surprisingly efficient training strategy that allows for improving model generalization and overall capability of the standard deep networks, by leveraging highly noisy labels. We observer that training a CNN from scratch using both clean and noisy data is better than just using the clean one. The contributions of this work are three-fold.

- We propose a CurriculumNet by developing en efficient learning strategy with curriculum learning. This allows us to train high-performance CNN models from large-scale web images with massive noisy labels, which are obtained without any human annotation.
- We design a new learning curriculum by ranking data complexity using distribution density in an unsupervised manner. This allows for an efficient implementation of curriculum learning tailored for this task, by directly exploring highly noisy labels.
- We conduct extensive experiments on a number of benchmarks, including WebVision [19], ImageNet [5], Clothing1M [39] and Food101 [2], where the proposed CurriculumNet obtains the state-of-the-art performance. The CurriculumNet, with an ensemble of multiple models, archives the top performance with a top-5 error rate of 5.2%, on WebVision Challenge at CVPR 2017, outperforming the other results by a large margin.

2 Related work

We give a brief review on recent studies developed for dealing with noisy annotations on image classification. For a comprehensive overview of label noise taxonomy and noise robust algorithms we refer to [7].

Recent approaches to learn from noisy web data can be roughly classified into two categories. (1) Methods aim to directly learn from noisy labels. This group of approaches mainly focus on noise-robust algorithms [16, 39, 25], and label cleansing methods which aim to remove or correct mislabeled data [3, 15]. However, they generally suffer from the main challenge of identifying mislabeled samples from hard training samples, which are crucial to improve model capability. (2) Semi-supervised learning approaches have also been developed to handle these shortcomings, by combining the noisy labels with a small set of clean labels [40, 26, 38]. A transfer learning approach solves the label noise by transferring

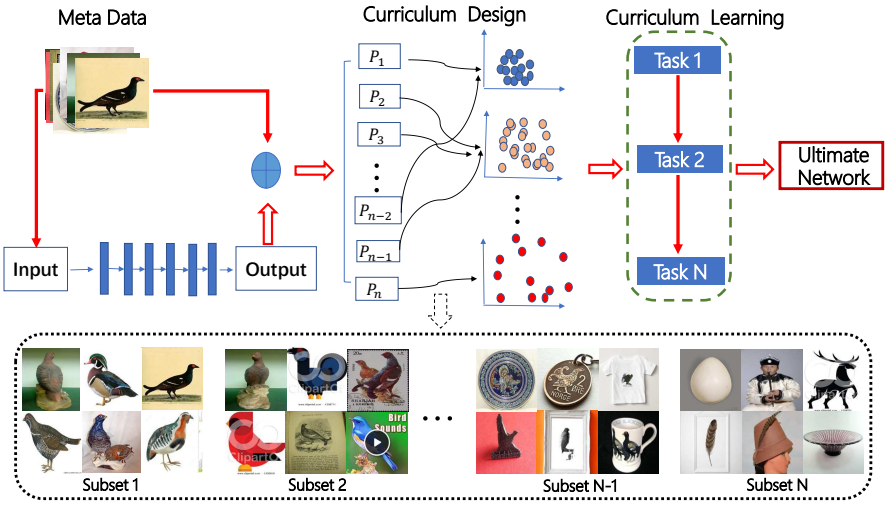

Fig. 2. Pipeline of the proposed WebNet. The training process includes three main steps initial features generation, curriculum design and curriculum learning.

correctness of labels to other classes [17]. The models trained on this subset are generalized to a larger dataset with unlabelled or weakly-labelled data [36]. Unlike these approaches, we do not propose a noise-cleansing or noise-robust or semi-supervised algorithm. Instead, we investigate improving model capability of the standard neural networks, by introducing a new training strategy that alleviates negative impact of the noisy labels.

Convolutional neural networks have recently been applied to training a robust model with noise data [39, 30, 25, 17, 15]. Xiao *et al.* [39] introduced a general framework to train CNNs with a limited number of human annotation, together with millions of noisy data. A behavior of CNNs on the training set with highly noisy labels was studied in [30]. MentorNet [15] improved the performance of CNNs trained on noise data, by learning an additional network that weights the training examples. Our method differs from these approaches by directly considering the mislabelled samples in our training process, and we show by experiments that with an efficient training scheme, a standard deep network is strongly robust against the highly noisy labels.

Our work is closely related to the work of [13], which is able to model noise arising from missing, but visually present labels. The method in [13] is conditioned on the input image, and was designed for multiple labels per image. It does not take the advantages of cleaned labels, and the focus is on missing labels, while our approach works reliably on the highly noisy labels, without any cleaned (manually annotated) one, and a learning curriculum is designed properly in a completely unsupervised manner.

3 Methodology

In this section, we present details of the proposed CurriculumNet motivated by human learning, in which the model starts from learning easier aspects of a concept, and then gradually take more complicated tasks into learning process [1]. We introduce a new method to design a learning curriculum in an unsupervised manner. Then CNNs are trained by following the designed curriculum, where the amount of noisy labels is increased gradually.

3.1 Overview

Pipeline of CurriculumNet is described in Fig. 2. It contains three main steps: (i) initial features generation, (ii) curriculum design and (iii) curriculum learning. First, we use all training data to learn an initial model which is then applied to computing a deep representation (e.g., fully-convolutional (fc) features) from each image in the training set. Second, the initial model aims to roughly map all training images into a feature space where the underlying structure and relationship of the images in each category can be discovered, providing an efficient approach that defines the complexity of the images. We explore the defined complexity to design a learning curriculum where all images in each category are split into a number of subsets ordered by complexity. Third, based on the designed curriculum, we employ curriculum learning which starts training CNNs from a easy subset which combines the easy subsets over all categories. It is assumed to have more clean images with correct labels in the easy subset. Then the model capability is improved gradually by continuously adding the data with increasing complexity into the training process.

3.2 Curriculum Design

Curriculum learning was originally proposed in [1]. It was recently applied to dealing with noise and outliers. One of the main issues to deliver advances of this learning idea is to design an efficient learning curriculum that is specific for our task. The designed curriculum should be able to discover meaningful underlying local structure of the large-scale noisy data in a particular feature space, and our goal is to design a learning curriculum able to rank the training images from easy to complex in an unsupervised manner. We apply a density based clustering algorithm that measures the complexity of training samples using data distribution density. Unlike previous approaches which were developed to handle noisy labels in small-scale or moderate-scale datasets, we design a new learning curriculum that allows our training strategy with a standard CNN to work practically well on the large-scale dataset, e.g., the WebVision database which contains over 2,400,000 web images with massive noisy labels.

Specifically, we aim to split the whole training set into a number of subsets, which are ranked from an easy subset having clean images with more reliable labels, to a more complex subset containing massive noisy labels. Inspired by recent clustering algorithm described in [29], we conduct following procedures

in each category. First, we train an initial model from the whole training set by using an Inception_v2 architecture [14]. Then all images in each category are projected into a deep feature space, by using the *fc*-layer features of the initial model, $P_i \rightarrow f(P_i)$ for each image P_i. Then we calculate an Euclidean distance matrix $D \subseteq \mathbb{R}^{n \times n}$ as,

$$D_{ij} = \|f(P_i) - f(P_j)\|^2 \tag{1}$$

where n is the number of images in current category, and D_{ij} indicates a similarity value between P_i and P_j(A smaller D_{ij} means higher similarity between P_i and P_j).

We first calculate a local density (ρ_i) for each image:

$$\rho_i = \sum_j X(D_{ij} - d_c) \tag{2}$$

where

$$X(d) = \begin{cases} 1 & d < 0 \\ 0 & \text{other} \end{cases}$$

where d_c is determined by sorting n^2 distances in $D \subseteq \mathbb{R}^{n \times n}$ from small values to large ones, and select a number which is ranked at $k\%$. This result is insensitive to the value of k between 50 and 70, and we empirically set $k = 60$ in all our experiments. ρ_i is the number of samples whose distances to i is smaller than d_c. It is natural to assume that a group of clean images with correct labels often have relatively similar visual appearance, and these images are projected closely to each other, leading to a large value of local density. By contrast, noisy images often have a significant visual diversity, resulting in a sparse distribution with a smaller value of the density.

Then we define a distance (δ_i) for each image:

$$\delta_i = \begin{cases} min_{j:\rho_j > \rho_i}(D_{ij}) & if\ \exists j\ s.t.\ \rho_j > \rho_i \\ max(D_{ij}) & \text{otherwise} \end{cases} \tag{3}$$

If there exist an image I_j having $\rho_j > \rho_i$, δ_i is $D_{i\hat{j}}$ where $\hat{j}$ is the sample nearest to i among the data. Otherwise, if δ_i is the largest one among all density, ρ_j is the distance between i and the data point which is farthest from i. Then a data point with the highest local density has the maximum value of δ, and is selected as cluster center for this category.

As we have computed a cluster center for the category, a closer data point to the cluster center, has a higher confidence to have a correct label. Therefore, we simply proceed k-mean algorithm to divide data points into a number of clusters, according to their distances to the cluster center, D_{cj}, where c is the cluster center. Fig. 3 (left) is an $\delta - \rho$ figure for all images in the category of cat from the WebVision dataset.

We generate three clusters in each category, and simply use the images within each cluster as a data subset. As each cluster has a density value measuring data distribution within it, and relationship between different clusters. This provides

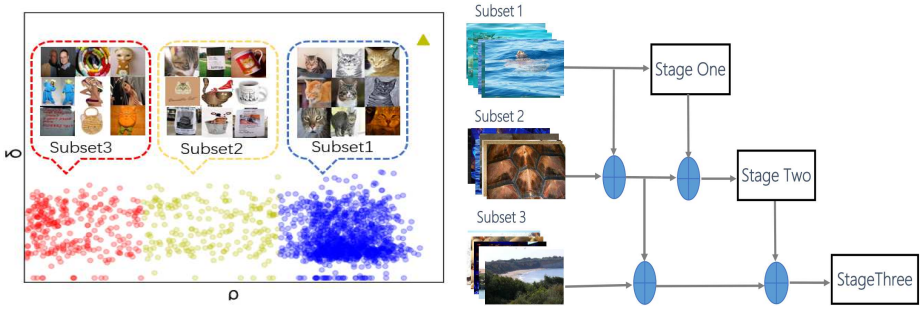

Fig. 3. Left: the sample of the cat category with three subsets. Right: learning process with designed curriculum.

a natural way to define the complexity of the subsets, giving a simple rule for designing a learning curriculum. A subset with a high density value means all images are closed to each other in the feature space, suggesting that these images have a strong similarity. We define this subset as a *clean* one, by assuming most of the labels are correct. The subset with a small density value means the images have a large diversity in visual appearance, which may include more irrelevant images with incorrect labels. This subset is considered as *noisy* data. Therefore, we generate a number of subsets in each category, arranged from clean, noisy, to highly noisy ones, which are ordered with increasing complexity. Each category has a same number of the subsets, and we combine them over all categories, which form our final learning curriculum that implements training sequentially on the clean, noisy and highly noisy subsets. Fig. 3 (left) show data distribution of the three subsets in the category of "cat" from the WebVision dataset, with a number of sample images. As can be found, images from the clean subset have very closed visual appearance, while the highly noisy subset contains a number of random images which are completely different from those in the clean subset.

3.3 Curriculum Learning

Learning process is performed by following the nature of data structure. The designed curriculum is able to discover underlying data structure based on visual appearance, in an unsupervised manner. We design a learning strategy which relies on the intuition - tasks are ordered by increasing difficulty, and training is proceeded sequentially from easier tasks to harder ones. We develop a multi-stage learning process that trains a standard neural network more efficiently with the enhanced capability for handling massive noisy labels.

Training details are described in Fig. 3 (right), where a convolutional model is trained through three stages by continuously mixing training subsets from clean subset to highly noisy one. Firstly, a standard convolutional architecture, such as Inception_v2 [14], is used. The model is trained by only using the clean data, where images within each category have close visual appearance. This allows

the model to learn basic but clear visual information from each category, severing as the fundamental features for the following process. Secondly, when the model trained in the first stage is convergence, we continue the learning process by adding the noise data, where images have more significant visual diversity, allowing the model to learn more meaningful and discriminative features from harder samples. Although the noise data may include incorrect labels, but it roughly preserves the main structure of the data, and thus leads to performance improvement. Thirdly, the model is further trained by adding the highly noise data which contains a large number of visual irrelevant images with incorrect labels. The deep features learned by following first two-stage curriculum are able to capture the main underlying structure of data. We observe that the highly noisy data added in the last stage does not impact negatively to the learned data structure. By contrast, it improves the generalization capability of the model, and allows the model to avoid over-fitting over the clean data, by providing a manner of regularization. A final model is obtained when the training is convergence in the last stage, where the three subsets are all combined. In addition, when samples from different subsets are combined in the second and third stages, we set different loss weights to the training samples of different subsets as 1, 0.5 and 0.5 for the clean, noisy and highly noisy subsets, respectively.

3.4 Implementation Details

Training Details: The scale of WebVision data [19] is significantly larger than that of ImageNet [5], it is important to considering the computational cost when extensive experiments are conducted in evaluation and comparisons. In our experiments, we employ the inception architecture with batch normalization (bn-inception) [14] as our standard architecture. The bn-inception model is trained by adopting the proposed density-ranking curriculum leaning. The network weights are optimized with mini-batch stochastic gradient decent (SGD), where the batch size is set to 256, and Root Mean Square Propagation (RMSprop) algorithm [14] is adopted. The learning rate starts from 0.1, and decreases by a factor of 10 at the iterations of $30 \times 10^4, 50 \times 10^4, 60 \times 10^4, 65 \times 10^4, 70 \times 10^4$. The whole training process stop at 70×10^4 iterations. To reduce the risk of over-fitting, we use common data augmentation technologies which include random cropping , scale jittering, and ratio jittering. We also add a dropout operation with a ratio of 0.2 after the global pooling layer.

Selective Data Balance: By comparing with ImageNet, another challenge of the WebVision data [18] is that the training images in different categories are highly unbalanced. For example, a large-scale category can have over 10,000 images, while a small-scale category only contains less than 400 images. CNN models, directly trained with random sampling on such unbalanced classes, will have a bias towards the large categories. To alleviate this problem, we develop a two-level data balance approach: subset-level balance and category-level balance. In the subset-level balance, training samples are selected in each min-batch as follows: (256, 0, 0), (128, 128, 0) and (128, 64, 64) for stage 1-3, respectively.

For the category-level balance, in each min-batch, we first random select 256 (in stage 1) or 128 (in stage 2 and 3) categories from the 1000 classes, and then we randomly select only one sample from each selected category. Notice that the category-level balance is only implemented on the clean subset. The performance was dropped down when we applied it to the noisy or highly noisy subset. Because we randomly collect a single sample from each category in the category-level balance, it is possible to obtain a single but complectly irrelevant sample from the noisy or highly noisy subset, which would negatively affect the training.

Multi-scale convolutional kernels: We also apply multi-scale convolutional kernels in the first convolutional layer, with three different kernel sizes: 5×5, 7×7 and 9×9. Then we concatenate three convolutional maps generated by three types of filters, which form the final feature maps of the first convolutional layer. The multi-scale filters enhance the low-level features in the first layer, leading to about 0.5% performance improvements on top-5 errors on the WebVision data.

4 Experimental Results and Comparisons

The proposed CurriculumNet is evaluated on four benchmarks: WebVision [19], ImageNet [5], Clothing1M [39] and Food101 [2]. Particularly, we investigate the learning capability on large-scale web images without human annotation.

4.1 Datasets

WebVision dataset [19] is an object-centric dataset, and is larger than ImageNet [5] for object recognition and classification. The images are crawled from both Flickr and Google images search, by using queries generated from the 1, 000 semantic concepts of the ILSVRC 2012. Meta information along with those web images (e.g., title, description, tags, etc.) are also crawled. The dataset for the WebVision 2017 contains 1,000 object categories (the same with the ImageNet). The training data contains 2,439,574 images in total, but without any human annotation. It includes massive noisy labels, as shown in Fig. 1. There are 50,000 manually-labeled images are used as validation set, and another 50,000 manually-labeled images for testing. The evaluation measure is based on top-5 error, where each algorithm provides a list of at most 5 object categories to match the ground truth.

Clothing1M dataset [39] is a large-scale fashion dataset, which includes 14 clothes categories. It contains 1 million noise label images and 74,000 manually annotated images. We call the annotated images as clean set, which is divided into training data,validation data and testing data, with numbers of 50,000,14,000, and 10,000 images, respectively. There are some images overlap between the clean set and the noisy set. The dataset was designed for learning robust models from noisy data without human supervision.

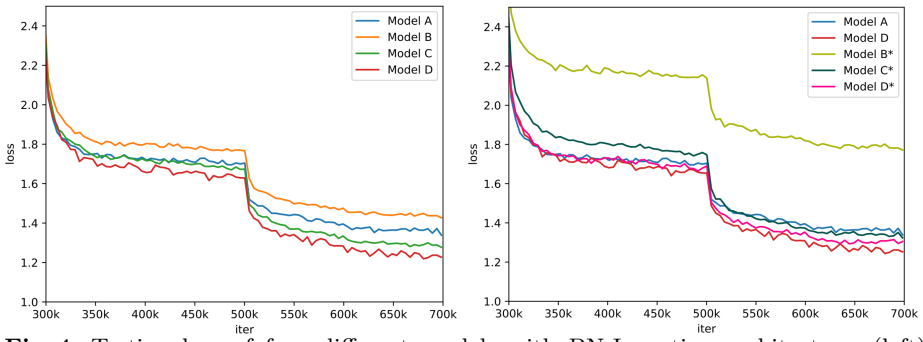

Fig. 4. Testing loss of four different models with BN-Inception architecture. (left) Density-based curriculum, and (right) K-mean based curriculum.

Food-101 dataset [2] is a standard benchmark to evaluate recognition accuracy of visual food. It contains 101 classes, with 101,000 real-world food images in total. The numbers of training and testing images are 750 and 250 per category, respectively. This is a clean dataset with full manual annotations provided. To conduct experiments with noise data, we manually add 20% noise images into the training set, which are randomly collected from the training set of ImageNet [5], and each image is randomly assigned a label from 101 categories from the Food-101.

4.2 Experiments and Comparisons

We conducted extensive experiments to evaluate the efficiency of the proposed approaches. we compare various training schemes by using the BN-Inception.

On training strategy. We evaluate four different training strategies by using a standard Inception_v2 architecture, resulting in four models, which are described as follows.

- **Model-A**: the model is trained by directly using the whole training set.
- **Model-B**: the model is trained by only using the clean subset.
- **Model-C**: the model is trained by using the proposed learning strategy, with a 2-subset curriculum: clean and noisy subsets.
- **Model-D**: the model is trained by using the proposed learning strategy, with a 3-subset curriculum: clean, noisy and highly noisy subsets.

Test loss of four models (on the validation set of the WebVision) are compared in Fig. 4, where the proposed CurriculumNet with a 2-subset curriculum and a 3-subset curriculum (Model-C and Model-D) have better convergence rates. Top 1 and Top 5 results of four models on the validation set of WebVision are reported in Table 1. The results are mainly consistent with the test loss presented in Fig. 4. The proposed method, with 3-subset curriculum learning, significantly outperforms the model trained on all data, with improvements of

Table 1. Top-1 and Top-5 errors (%) of four different models with BN-Inception architecture on validation set. The models are trained on Webvision training set and tested on WebVision and ILSVRC validation sets under various models

Mothed	**WebVision**		**ImageNet**	
	Top-1	Top-5	Top-1	Top-5
Model-A	30.16	12.43	36.00	16.20
Model-B	30.28	12.98	37.09	16.42
Model-C	28.44	11.38	35.66	15.24
Model-D	**27.91**	**10.82**	**35.24**	**15.11**

30.16% $\rightarrow$ 27.91% and 12.43% $\rightarrow$ 10.82% on Top 1 and Top 5 errors, respectively. These improvements are significant on such a large-scale challenge. Consistent improvements are obtained on the validation set of ImageNet, where the models were trained on the WebVision data. In all 1000 categories, our approaches lead to performance improvements on 668 categories, while only 195 categories reduced their Top 5 results, and the results of remained 137 categories were unchanged.

On highly noisy data or training labels. We further investigate the impact of highly noisy data to the proposed learning strategy. We used different percentages of data from the highly noisy subset for 3-subset curriculum learning, ranging from 0% to 100%. Results are reported in Table 2. As shown, the best results on both Top 1 and Top 5 are achieved at 50% of the highly noisy data. This suggests that, by using the proposed training method, even the highly noisy data can improve model generalization capability, by increasing the amount of the training data with more significant diversity, demonstrating the efficiency of the proposed approach. Increasing the amount of highly noisy data further did not improvement the performance, but with very limited negative affect.

To provide more insights and give deeper analysis on label noise exist, we applied most recent ImageNet-trained SEnet [12] (which has a Top 5 error of 4.47% on ImageNet) to classify all images from the training set of the WebVision data. *We assume the output label of each image by SEnet is correct, and compute the rate of correct labels in each category.* We observed that the average noise rate over the whole training set of the WebVision data is high to 52% (Top 1), indicating that a large amount of incorrect labels is included. We further compute the average noise rates for three subsets of the designed learning curriculum, which are 65%, 39% and 15%, respectively. These numbers are consistent with the increasing complexity of the three subsets, and suggest that most of images in the third subset are highly noisy.

We calculate the number of categories in 10 different intervals of the correct rates of the training labels, as shown in Fig. 5 (left). There are 12 categories having a correct rate that is lower than 10%. We further compute the average performance gain in each interval, as show in Fig. 5 (right). We found that the categories with lower correct rates (e.g., $< 40\%$) have larger performance

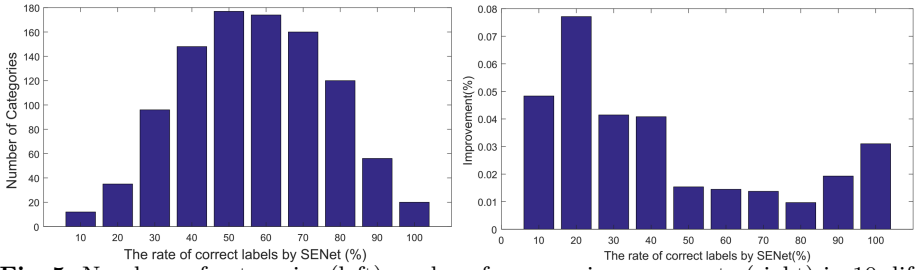

Fig. 5. Numbers of categories (left), and performance improvements (right) in 10 different rate intervals of the training labels.

gains ($> 4\%$), and the most significant improvement happens in the interval of 10%-20%, which has an improvement of 7.7%.

On different clustering algorithms. The proposed clustering based curriculum learning can generalize well to other clustering algorithms. We verify it by comparing our density based curriculum design with K-mean based clustering on the proposed 3-subset CurriculumNet. As shown in Fig. 4 (right), the Model-B* which is trained using the clean subset by K-mean has a significantly lower performance, which means that training without the proposed curriculum learning is highly sensitive to the quality. By adapting proposed method, Model-D* significantly improves the performance, from 16.6% to 11.5% (Top 5), which is comparable to Model-D. These results demonstrate the strong robustness of the proposed CurriculumNet, allowing for various qualities of the data generated by different algorithms.

Final results on the WebVision challenge. We further evaluate the performance of CurriculumNet (Model-D) by using various networks architectures, including Inception_v2 [14], Inception_v3 [35], Inception_v4 [33] and Inception_resnet_v2 [33]. Results are reported in Table 3. As can be found, the Inception_v3 outperforms the Inception_v2 substantially, from 10.82% to 7.88% on the Top 5, while a more complicated model, such as Inception_v4 and Inception_resnet_v2, only has similar performance with a marginal performance gain obtained.

Our final results were obtained with ensemble of six models. We had the best performance at a Top 5 error of 5.2% on the WebVision challenge 2017[18]. It outperforms the 2nd one by a margin of about 2.5%, which is about 50% relative error, and thus is significant for this challenging task. The 5.2% Top 5 error is also comparable to human performance on the ImageNet, but our methods obtained this result by using weakly-supervised training data without any human annotation.

Comparisons with the state-of-the-art methods. Our method is evaluated by comparing it with recent state-of-the-art approaches developed specifically for learning from label noise, such as CleanNet [17], FoodNet [24] and Patrini *et.*

Table 2. Performance (%) of model-D by using various percentages of data from the highly noisy subset.

Noise data(%)	Top1	Top5
0	28.44	11.38
25%	28.17	10.93
50%	**27.91**	**10.82**
75%	28.48	11.07
100%	28.33	10.94

Table 3. Performance (%) of model-D by using various networks.

Networks	Top1	Top5
Inception_v2	27.91	10.82
Inception_v3	22.21	7.88
Inception_v4	21.97	6.64
Inception_resnet_v2	**20.70**	**6.38**

al.'s approach [25]. Experiments and comparisons are conducted on four benchmarks: WebVision [19], ImageNet [5], Clothing1M [39] and Food101 [2]. Model-D with Inception_v2 is used in all our experiments. By following [17], we use the training set of WebVision to train the models, and test on the validation sets of the WebVision and ILSVRC, both of which has same 1000 categories. On the Clothing1M, we conduct two groups of experiments by following [17], we first apply our curriculum-based training method to one million noisy data, and then use 50K clean data to fine-tune the trained model. We compare both results against CleanNet [17] and the approach of Patrini *et. al.* [25].

Full results are presented in Table 4. CurriculumNet improves the performance of our baseline significantly in all four databases. Furthermore, our results compare favorably against recent CleanNet on all datasets, with consistent improvements ranged from about 1.5% to 3.3%. Particularly, CurriculumNet reduces Top 5 error of the CleanNet from 12.2% to 10.8% on the WebVision data. In addition, CurriculumNet also outperforms Patrini *et. al.*'s approach (19.6%→18.5%) [25] on the Clothing1M. On the Food101, CurriculumNet, trained with 20% additional noise data with *completely random labels*, achieved substantial improvements over both CleanNet (16.0%→12.7%) and FoodNet (27.9%→12.7%) [24]. These remarkable improvements confirm the advances of CurriculumNet, demonstrating strong capability for learning from massive amount of noisy labels.

Train with more clean data: WebVision+ImageNet. We evaluate the performance of CurriculumNet by increasing the amount of clean data in the training set of WebVision. Since ImageNet data is fully cleaned and manually annotated, a straightforward approach is to simply combine the training sets of WebVision and ImageNet data. We implement CurriculumNet with Inception_v2 by considering ImageNet data as an additional clean subset, and test the results on the validation sets of both databases. Results are reported in Table 5.

We summary key observations as follows. (i) By combining WebVision data into ImageNet data, the performance is generally improved due to the increased amount of training data. (ii) Performance of the proposed CurriculumNet is improved significantly on both validation sets by increasing the amount of the clean data (ImageNet), such as 10.8%→8.5% on WebVision, and 15.1%→7.1%

Table 4. Comparisons with most recent results on the Webvision, ImageNet, Clothes-1M and Food101 databases. For the Webvision and ImageNet, the models are trained on WebVision training set and tested on WebVision and ILSVRC validation sets.

Mothed	**WebVision**	**ImageNet**	**Clothing1M**	**Food101**
	Top-1(Top-5)	Top-1(Top-5)	Top-1	Top-1
Baseline[17]	32.2(14.2)	41.1(20.2)	24.8	18.3
CleanNet [17]	29.7(12.2)	36.6(15.4)	20.1	16.0
MentorNet [15]	29.2(12.0)	37.5(17.0)	–	–
Our Baseline	30.3(13.0)	37.1(16.4)	24.2	15.0
CurriculumNet	**27.9(10.8)**	**35.2(15.1)**	**18.5**	**12.7**

Table 5. Performance on the validation sets of ImageNet and WebVision. Models are trained on the training set of ImageNet, WebVision or ImageNet+WebVision.

Traning Data	**WebVision**		**ImageNet**	
	Top-1	Top-5	Top-1	Top-5
ImageNet	32.8	13.9	26.9	8.6
ImageNet+WebVision	25.3	9.0	25.6	7.4
CurriculumNet(WebVision)	27.9	10.8	35.2	15.1
CurriculumNet(WebVision+ImageNet)	**24.7**	**8.5**	**24.8**	**7.1**

on ImageNet. (iii) By using both WebVision and ImageNet as training data, CurriculumNet is able to improve the performance on both validation sets. For example, it reduces the Top 5 error of WebVision from 9.0% to 8.5% with a same training set. (iv) On ImageNet, CurriculumNet boosts the performance from a Top 5 error of 8.6% to 7.1%, by leveraging additional noisy data (e.g., WebVision). This performance gain is significant on ImageNet, which further confirms the strong capability of CurriculumNet on learning from noisy data.

5 Conclusion

We have presented a CurriculumNet - an new training strategy able to train CNN models more efficiently on large-scale weakly-supervised web images, where no human annotation is provided. By leveraging the idea of curriculum learning, we propose a novel learning curriculum by measuring data complexity using cluster density. We show by experiments that the proposed approaches have strong capability for dealing with massive noisy labels. They not only reduce the negative affect of noisy labels, but also, notably, improve the model generalization ability by using the highly noisy data. The proposed CurriculumNet achieved the state-of-the-art performance on the Webvision, ImageNet, Clothing-1M and Food-101 benchmarks. With an ensemble of multiple models, it obtained a Top 5 error of 5.2% on the Webvision Challenge 2017, which outperforms the other submissions by a large margin of about 50% relative error rate.

References

1. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: ICML. pp. 41–48. ACM (2009)
2. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: ECCV. pp. 446–461. Springer (2014)
3. Brodley, C.E., Friedl, M.A.: Identifying mislabeled training data. CoRR, abs/1106.0219 (1999)
4. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. CoRR **abs/1606.00915** (2016)
5. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F.: ImageNet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009)
6. Everingham, M., Gool, L.V., Williams, C., Winn, J., Zisserman, A.: The pascal visual object classes challenge 2007 (voc 2007) results (2007). In: URL http://www. pascal-network. org/challenges/VOC/voc2007/workshop/index. html (2008)
7. Frénay, B., Verleysen, M.: Classification in the presence of label noise: a survey. IEEE transactions on neural networks and learning systems **25**(5), 845–869 (2014)
8. Guo, S., Huang, W., Wang, L., Qiao, Y.: Locally-supervised deep hybrid model for scene recognition. IEEE Trans. on Image Processing (TIP) **26**, 808–820 (2017)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2016), cVPR
10. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV. pp. 2980–2988 (2017)
11. Hong, S., Noh, H., Han, B.: Decoupled deep neural network for semi-supervised semantic segmentation. In: NIPS. pp. 1495–1503 (2015)
12. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. CVPR (2018)
13. I. Misra, C. L. Zitnick, M.M., Girshick, R.: Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In: CVPR (2016)
14. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. CoRR **abs/1502.03167** (2015)
15. Jiang, L., Zhou, Z., Leung, T., Li, L.J., Fei-Fei, L.: Mentornet: Regularizing very deep neural networks on corrupted labels. CoRR **abs/1712.05055** (2017)
16. Larsen, J., Nonboe, L., Hintz-Madsen, M., Hansen, L.K.: Design of robust neural network classifiers. In: ICASSP (1998)
17. Lee, K.H., He, X., Zhang, L., Yang, L.: Cleannet: Transfer learning for scalable image classifier training with label noise. CoRR **abs/1711.07131** (2017)
18. Li, W., Wang, L., Li, W., Agustsson, E., Berent, J., Gupta, A., Sukthankar, R., Van Gool, L.: Webvision challenge: Visual learning and understanding with web data. CoRR **abs/1705.05640** (2017)
19. Li, W., Wang, L., Li, W., Agustsson, E., Van Gool, L.: Webvision database: Visual learning and understanding from web data. CoRR **abs/1708.02862** (2017)
20. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection pp. 2980–2988 (2017)
21. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755 (2014)
22. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: ECCV. pp. 21–37. Springer (2016)

23. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. pp. 3431–3440 (2015)
24. Pandey, P., Deepthi, A., Mandal, B., Puhan, N.: Foodnet: Recognizing foods using ensemble of deep networks. IEEE Signal Processing Letters **24**(12), 1758–1762 (2017)
25. Patrini, G., Rozza, A., Menon, A.K., Nock, R., Qu, L.: Making deep neural networks robust to label noise: a loss correction approach pp. 1944–1952 (2017)
26. R. Fergus, Y.W., Torralba, A.: Semi-supervised learning in gigantic image collections. In: NIPS (2009)
27. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR. pp. 779–788 (2016)
28. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NIPS. pp. 91–99 (2015)
29. Rodriguez, A., Laio, A.: Clustering by fast search and find of density peaks. Science **344**(6191), 1492–1496 (2014)
30. Rolnick, D., Veit, A., Belongie, S., Shavit, N.: Deep learning is robust to massive label noise. CoRR **abs/1705.10694** (2017)
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2014)
32. Sukhbaatar, S., Fergus, R.: Learning from noisy labels with deep neural networks. CoRR **abs/1406.2080** (2014)
33. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI. pp. 4278–4284 (2017)
34. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR. pp. 1–9 (2015)
35. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR. pp. 2818–2826 (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 (2017)
37. Wang, L., Guo, S., Huang, W., Xiong, Y., Qiao, Y.: Knowledge guided disambiguation for large-scale scene classification with multi-resolution cnns. IEEE Trans. on Image Processing (TIP) **26**, 2055–2068 (2017)
38. X. Chen, A.S., Neil, A.G.: Extracting visual knowledge from web data. In: ICCV (2013)
39. Xiao, T., Xia, T., Yang, Y., Huang, C., Wang, X.: Learning from massive noisy labeled data for image classification. In: CVPR. pp. 2691–2699 (2015)
40. Zhu, X.: Semi-supervised learning literature survey. CoRR, abs/1106.0219 (2005)