Web-Supervised Network with Softly Update-Drop Training for Fine-Grained Visual Classification

Chuanyi Zhang,[1] Yazhou Yao,[1*] Huafeng Liu,[1] Guo-Sen Xie,[2] Xiangbo Shu,[1]
Tianfei Zhou,[2] Zheng Zhang,[3] Fumin Shen,[4] Zhenmin Tang[1]
[1]Nanjing University of Science and Technology, China, [2]Inception Institute of Artificial Intelligence, UAE
[3]Harbin Institute of Technology, Shenzhen, China, [4]University of Electronic Science and Technology of China

Abstract

Labeling objects at the subordinate level typically requires expert knowledge, which is not always available from a random annotator. Accordingly, learning directly from web images for fine-grained visual classification (FGVC) has attracted broad attention. However, the existence of noise in web images is a huge obstacle for training robust deep neural networks. In this paper, we propose a novel approach to remove irrelevant samples from the real-world web images during training, and only utilize useful images for updating the networks. Thus, our network can alleviate the harmful effects caused by irrelevant noisy web images to achieve better performance. Extensive experiments on three commonly used fine-grained datasets demonstrate that our approach is much superior to state-of-the-art webly supervised methods. The data and source code of this work have been made anonymously available at: https://github.com/z337-408/WSNFGVC.

Introduction

DNNs have achieved impressive results on many computer vision tasks due to available large-scale image datasets (Deng et al. 2009; Bai et al. 2018a; 2018b). However, fine-grained image classification remains challenging. Subdividing a category into subcategories multiplies the number of labels and thus is a labor-intensive and time-consuming problem. What's worse, fine-grained annotation usually requires expert knowledge, which exacerbates the labeling problem. To reduce the cost of manual labeling, some works focused on the semi-supervised paradigm (Xu et al. 2015; Cui et al. 2016; Niu, Veeraraghavan, and Sabharwal 2018). However, these works inevitably involve various forms of human intervention and remain labor-consuming.

Compared to manual-labeled image datasets, web images are a rich and free resource. For arbitrary categories, the potential training data can be easily obtained from the image search engines like Google or Bing. Therefore, it is a natural idea to directly leverage web images for training fine-grained classification models. Unfortunately, due to the error-index of image search engine, the precision of returned images from image search engine is still unsatisfactory. As pointed by (Schroff, Criminisi, and Zisserman 2011), the average precision of the top 1000 images for 18 categories from Google Image Search engine is only 32%.

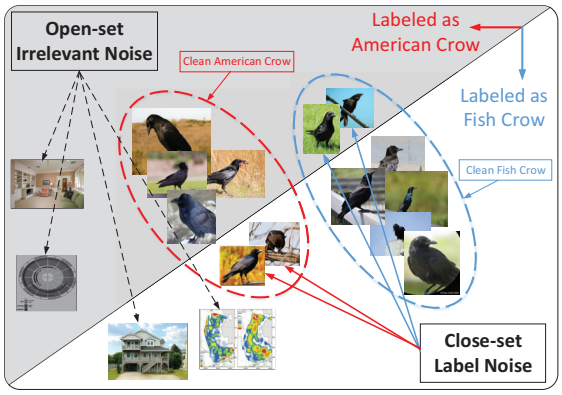

Figure 1: An illustration of the close-set and open-set noisy web images on a bird dataset. Noisy images in the close-set have their true labels in the dataset. Open-set contains irrelevant noisy images, whose labels are outside of the dataset.

As shown in Fig. 1, the noisy web images for fine-grained categories can be divided into two groups: Close-set noise and Open-set noise. Specifically, noisy images in the close-set have their true labels in the dataset (*e.g.*, the images of "Fish Crow" are mistakenly labeled as "American Crow" in Fig. 1). Open-set consists of irrelevant noisy images, whose labels are outside of the dataset (*e.g.*, the images pointed by the black dotted arrow in Fig. 1). As deep neural networks have a high capacity to fit noisy data (Arpit et al. 2017; Zhang et al. 2016), training fine-grained neural network models directly with these noisy web images will result in poor performance.

To reduce the harmful influence of noise, some works concentrate on estimating the noise transition probabilities between different category labels. For example, (Reed et al. 2014) leveraged bootstrapping loss and assigned a weight to the current prediction to compensate for the erroneous guiding of noisy samples. (Goldberger and Ben-Reuven 2016) added an additional softmax layer to estimate the label noise

*Corresponding author: Yazhou Yao

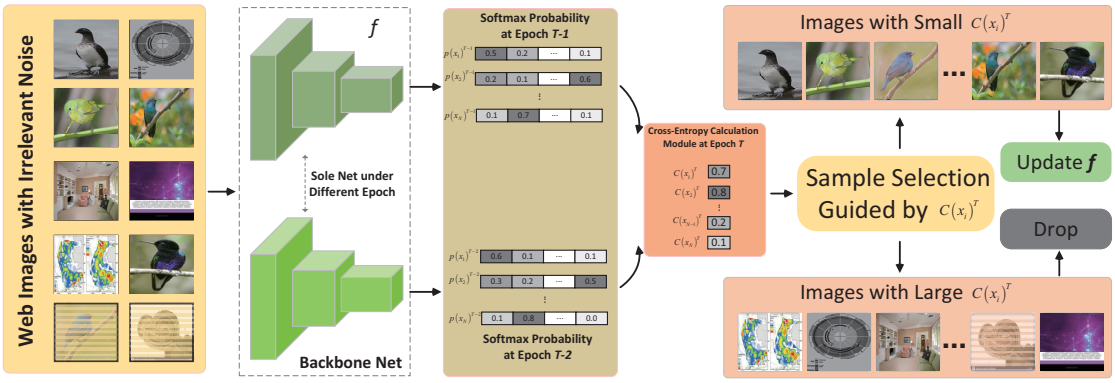

Figure 2: The architecture of our model. For each input web image x_i, we first obtain the softmax probability of epoch $T-1$, $T-2$ as $\boldsymbol{p}(x_i)^{T-1}$ and $\boldsymbol{p}(x_i)^{T-2}$, respectively. Then we compute the cross-entropy $C(x_i)^T$ between $\boldsymbol{p}(x_i)^{T-2}$ and $\boldsymbol{p}(x_i)^{T-1}$ in epoch T. $C(x_i)^T$ is leveraged to supervise the separation of useful and irrelevant noisy web images. To be specific, images with large $C(x_i)^T$ are identified as irrelevant noisy images and then dropped during training. Those with small $C(x_i)^T$ are regarded as useful images and utilized to further update the network $\boldsymbol{f}$.

transition matrix. Unfortunately, exact recovery of the noise transition probabilities is difficult and remains a challenging problem far from being solved. Alternatively, another branch of works endeavors to focus on sample section mechanism, which tries to separate clean samples from noise. The representative works are Decoupling (Malach and Shalev-Shwartz 2017) and Co-teaching (Han et al. 2018). These works identified clean samples out of the mini-batch and used them to update networks. Nevertheless, they assume a close-set noisy label setting, where all samples, including noisy ones, have their true labels in the dataset. Such restricted assumption contradicts the more practical open-set scenario. Ignoring the existence of irrelevant noise makes above-mentioned works less practical. As the web-supervised learning gets popular, some works focus on open-set scenario (Wang et al. 2018; Liang, Li, and Srikant 2017) to tackle the irrelevant noise. Unfortunately, neither of them is designed for fine-grained classification.

In this paper, we propose a simple yet effective sample selection approach to remove irrelevant noise from the training set for web-supervised fine-grained tasks. Our work is motivated by the following observations: 1) Soft labels contain more information than one-hot label (Hinton, Vinyals, and Dean 2015), especially in fine-grained classification tasks, where subcategories share obvious similarities. 2) Deep neural networks always memorize easy instances first, and gradually adapt to hard instances and noisy instances (Arpit et al. 2017; Zhang et al. 2016). 3) Selecting samples globally is more reliable than doing that in mini-batches.

During training, we aim to identify irrelevant noisy samples and drop them. Unlike most existing methods which use the loss to find noisy samples, we leverage the cross-entropy of the softmax probability between the contiguous epochs instead (we name it as probability cross-entropy in the following). The proposed approach can make good use of the information encoded in the soft label and is able to measure the prediction changes of the network. In addition, open-set irrelevant noisy samples are harder to fit than clean samples. The predictions of them are unstable and changing rapidly during training, resulting in big probability cross-entropy. Thus, irrelevant noisy samples can be distinguished from the useful training set and dropped by calculating the probability cross-entropy. In this way, the network can alleviate the harmful effect of open-set noise and achieve better performance. Extensive experiments and ablation studies demonstrate that our approach outperforms state-of-the-art methods. Our learning paradigm delivers a new pipeline for fine-grained visual classification, which is more practical for real-world applications.

Related Work

Fine-grained Classification

The task of fine-grained classification is to distinguish objects at subordinate level. Due to the similarities between subcategories, the early works train the network to learn discriminative features by utilizing strong annotation like bounding boxes or part annotations (Wei et al. 2018; Huang et al. 2016; Zhang et al. 2014; Lin et al. 2015). Despite promising results, these strongly supervised methods demand heavy human annotation. To overcome this drawback, recent studies focus on weakly supervised methods, which only need image-level labels (Fu, Zheng, and Mei 2017; Lin, RoyChowdhury, and Maji 2015; Zheng et al. 2017; Wang, Morariu, and Davis 2018). To further improve the performance by using more training images, some web-supervised methods manage to leverage easily accessible web data (Niu, Veeraraghavan, and Sabharwal 2018; Cui et al. 2016; Xu et al. 2016; Niu, Li, and Xu 2015; Xiao et al. 2015; Krause et al. 2016). Our approach is a pure web-supervised method, only using web images and requiring no human intervention.

Web-supervised Learning

Because web images are abundant and easy to obtain, web-supervised learning is drawing more attention(Yao et al. 2019;

2018; 2017; 2018). However, noisy images in the web dataset are harmful. Directly training fine-grained models on web dataset tends to show poor performance. To overcome this problem, numerous studies have been performed, which can be categorized into two strands: loss correction approaches and sample selection methods. Existing loss correction approaches trying to estimate label noise transition like (Goldberger and Ben-Reuven 2016) are unable to cope with the open-set irrelevant noisy web images. This is because the true labels of irrelevant noisy samples are outside the set of training labels.

Conversely, it's not a problem for sample selection methods, because they take no account of the true labels of irrelevant noisy samples and just select clean instances out of the noisy ones. Owing to this advantage, sample selection methods are more practical in an open-set scenario. The representative sample selection methods contain Decoupling (Malach and Shalev-Shwartz 2017), Co-teaching (Han et al. 2018) and MentorNet (Jiang et al. 2017). Decoupling and Co-teaching both train two peer networks simultaneously. Specifically, Decoupling chooses samples that have different predictions from two networks as useful ones. Co-teaching let each network selects small-loss samples as clean ones for its peer network. MentorNet trains an additional teacher network to teach a student network by providing clean samples, whose labels are probably correct. However, these works are mainly designed for a manually noisy dataset, a close-set scenario, making them less practical in real-world noisy web images. To overcome this drawback, our work is designed for open-set scenario and has the ability to cope with irrelevant noisy samples in the real-world dataset.

The Proposed Approach

In this section, we present a simple yet effective training mechanism to learn robust web-supervised fine-grained models. As the training progresses, our training mechanism is able to dynamically drop irrelevant noisy images and purify the web training set, thus achieves better performance.

Our proposed framework is present in Fig. 2. Assume that the neural network $\boldsymbol{f} = (f_1, ..., f_M)$ is trained to classify M classes. At each epoch T, we first utilize the network output logits $\boldsymbol{f}(x_i)$ to compute the softmax probability $\boldsymbol{p}(x_i)^T = (p_1(x_i)^T, ..., p_M(x_i)^T)$ for each instance x_i in the training set $\mathcal{D}$:

$$p_j(x_i)^T = \frac{\exp(f_j(x_i))}{\sum_{s=1}^{M} \exp(f_s(x_i))}. \tag{1}$$

Then when epoch $T > 2$, for each instance x_i, we compute the softmax probability cross-entropy $C(x_i)^T$ between $\boldsymbol{p}(x_i)^{T-2}$ and $\boldsymbol{p}(x_i)^{T-1}$ through

$$C(x_i)^T = -\sum_{j=1}^{M} p_j(x_i)^{T-1} \log p_j(x_i)^{T-2}. \tag{2}$$

Next, we select samples from the whole training set and set images with small cross-entropy $C(x)^T$ as useful images. By doing this, we can form a selected set $\widehat{\mathcal{D}}^T$. Those with big cross-entropy $C(x)^T$ are regarded as irrelevant noisy images and won't be used for training. The number of selected samples is controlled by a drop rate

$$r(T) = \tau \cdot \min\{\frac{T}{T_k}, 1\}, \tag{3}$$

which is dynamically updated during training. $\widehat{\mathcal{D}}^T$ can be obtained by solving the following problem:

$$\widehat{\mathcal{D}}^T = \underset{\mathcal{D}':|\mathcal{D}'|\geq(1-r(T))|\mathcal{D}|}{\arg\min} \sum\nolimits_{x \in \mathcal{D}'} C(x)^T. \tag{4}$$

Finally, we only leverage selected set $\widehat{\mathcal{D}}^T$ to update the network $\boldsymbol{f}$. The detailed steps of our proposed approach can be summarized in the Algorithm 1.

Algorithm 1: Softly Update-Drop Training

Input: Initialized network $\boldsymbol{f}$, training set $\mathcal{D}$, maximum drop rate τ, epoch T_k and $T_{\max}$.
for $T = 1, 2, ..., T_{\max}$ **do**
  for each instance x_i in training set $\mathcal{D}$ **do**
    if $T > 2$ **then**
      Compute $C(x_i)^T$ according to Eq. (2)
      Obtain $\widehat{\mathcal{D}}^T$ according to Eq. (4)
    else
      $\widehat{\mathcal{D}}^T = \mathcal{D}$
    end
    Compute $\boldsymbol{p}(x_i)^T$ according to Eq. (1)
  end
  Update $\boldsymbol{f}$ and $r(T)$
end
Output: Updated network $\boldsymbol{f}$

Noise Identification

Memorization effects (Arpit et al. 2017; Zhang et al. 2016) indicate that deep neural networks always fit to easy examples in the initial epochs, and gradually adapt to hard examples and noisy examples. Moreover, the irrelevant noisy samples are totally different from clean ones in the training set, making them more difficult for the network to fit. Therefore, the prediction results of irrelevant noisy samples change rapidly during training, especially at the early stages. Accordingly, we can identify them by measuring how rapidly the predictions change.

To measure the prediction changes, we compute the cross-entropy of softmax probability between contiguous epochs. Predictions of irrelevant noisy samples change more rapidly than that of clean ones, resulting in higher probability cross-entropy. For useful samples x in the selected set $\widehat{\mathcal{D}}$ and noisy samples $\widetilde{x}$ in dropped set $(\mathcal{D} - \widehat{\mathcal{D}})$, we have

$$\frac{1}{|\mathcal{D} - \widehat{\mathcal{D}}|} \sum_{\widetilde{x} \in (\mathcal{D} - \widehat{\mathcal{D}})} C(\widetilde{x}) > \frac{1}{|\widehat{\mathcal{D}}|} \sum_{x \in (\widehat{\mathcal{D}})} C(x). \tag{5}$$

Then we can choose samples which have small probability cross-entropy $C(x)$ as useful images and use them to update the network. Different from existing methods that directly leverage cross-entropy (a hard target), our method utilizes the

Table 1: ACA (%) performances on three benchmark fine-grained datasets. BBox/Anno (✓) indicates human annotations are utilized during training. Training set shows whether the dataset is manually labeled (anno.) or collected from the web (web). iNat means the iNaturalist dataset.

	Method	Publication	BBox/Anno	Training Set	Datasets		
					CUB200	Aircrafts	Cars-196
α	Mask-CNN	PR 2018	✓	anno.	85.70	-	-
β	Bilinear CNN	ICCV 2015		anno.	84.10	83.90	91.30
	RA-CNN	CVPR 2017		anno.	85.30	-	92.50
	Multi-attention	ICCV 2017		anno.	86.50	89.90	92.80
	Filter-bank	CVPR 2018		anno.	86.70	92.00	93.80
γ	Xu *et al.*	TPAMI 2018	✓	anno.+web	84.60	-	-
	Cui *et al.*	CVPR 2016	✓	anno.+web	80.70	-	-
	Niu *et al.*	CVPR 2018		anno.+web	76.47	-	-
	Cui *et al.*	CVPR 2018		anno.+iNat	89.29	90.70	93.50
ζ	WSDG	CVPR 2015		web	70.61	-	-
	Xiao *et al.*	CVPR 2015		web	70.92	-	-
	Decoupling	NeurIPS 2017		web	70.56	75.97	75.00
	Co-teaching	NeurIPS 2018		web	73.85	72.76	73.10
η	**Ours**	-		web	**77.22**	**72.88**	**78.71**

α : strongly supervised β : weakly supervised γ : semi-supervised ζ : webly supervised η : Ours

probability, a soft label, to identify noisy examples. Softmax probability can better describe the predictions than one-hot output, thus the probability cross-entropy between contiguous epochs can better represent the prediction changes than the cross-entropy. By this way, we can distinguish irrelevant noise from web training set more efficiently than existing methods.

Global Sampling

Existing methods perform sample selection mainly through mini-batch (Song, Kim, and Lee 2019; Han et al. 2018). The number of noisy images N_i in a mini-batch i forms a hypergeometric distribution. Given the noise rate $R_{\mathcal{D}}$ of dataset $\mathcal{D}$ and batch size N_b, we have

$$N_i \sim H(|\mathcal{D}|, |\mathcal{D}| \cdot R_{\mathcal{D}}, N_b). \tag{6}$$

In this distribution, the number of noisy images N_i fluctuate among different batches, resulting in the noise rate imbalance problem. Some batches may have less noisy samples while others have more. In the case that drop rate $r(T)$ is fixed in each epoch, clean samples might have to be dropped in some mini-batches while noisy samples are used for training in other mini-batches. So the sample selection in mini-batches is unstable and unreliable. To overcome the noise rate imbalance problem, we choose to select samples from the whole training set. Through making the selection results more stable, better performance can be achieved.

Dynamic Drop Rate

Similar to Co-teaching (Han et al. 2018), we utilize a linear increased drop rate $r(T)$ in the early training epochs. As indicated by memorization effects (Arpit et al. 2017; Zhang et al. 2016), deep neural networks have the ability to filter out noisy instances using their loss values at the early training stage. Then deep neural networks will eventually overfit on noisy samples as the number of epochs increases. To leverage this property, we dynamically increase the drop rate $r(T)$ and manage to keep more instances at early epochs and increasingly drop noisy images before they are memorized.

Experiments

Datasets We evaluate our approach on three benchmark fine-grained datasets, CUB200-2011 (Wah et al. 2011), FGVC-aircraft (Maji et al. 2013), and Cars-196 (Krause et al. 2013). Average Classification Accuracy (**ACA**) is taken as the evaluation metric.

Implementation Details Following (Niu, Veeraraghavan, and Sabharwal 2018), we collect web images by using labels in benchmark datasets. We treat the retrieved web images as the training set and directly adopt the testing data from CUB200-2011, FGVC-aircraft, and Cars-196 as the test set to build three web-supervised datasets (Bird dataset, Aircraft dataset, and Car dataset). To ensure the collected web images have no overlap with the test images in CUB200-2011, FGVC-aircraft, and Cars-196, we perform PCA-based near-duplicate removal (Zhou et al. 2016) between retrieved web images and test images.

In the experiments, we use a pre-trained model, VGG-16 (Simonyan and Zisserman 2014), to initialize the network. We select the maximum drop rate τ from the values of $\{0.15, 0.2, 0.25, 0.3\}$ and epoch T_k from the values of $\{5, 10, 15, 20\}$. Through experiments, we ultimately set $\tau = 0.25$ and $T_k = 10$ as the default value on CUB200 and FGVC-aircraft datasets, and set $\tau = 0.20$ and $T_k = 10$ on Cars-196 dataset. For model training, we follow (Lin, RoyChowdhury, and Maji 2015) and adopt a two-step training strategy. Specifically, we first freeze the convolutional layer parameters and only optimize the last fully connected layer. Then we optimize all layers in the previously learned model. In the experiment, we use Adam optimizer with momentum=0.9. The learning rate, batch size, and epoch number in the first step are set to 0.001, 128 and 200, while in the second step, they are set to 0.0001, 64 and 100, respectively.

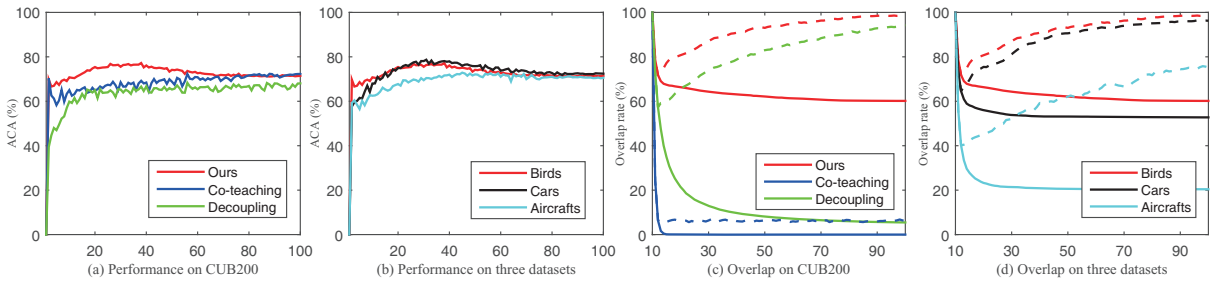

Figure 3: Test accuracy and Overlap rate vs. number of epochs. (a): Test accuracies of our approach, Co-teaching and Decoupling on CUB200; (b): Test accuracies of our approach on three benchmark datasets; (c): Overlap rates of our approach, Co-teaching and Decoupling on CUB200; (d): Overlap rates of our approach on three benchmark datasets. The overlap rate of all previous epochs is plotted with solid lines and the overlap rate of three contiguous epochs (*e.g.*, epoch T_{i-2}, T_{i-1} and T_i) is plotted with dotted lines in (c) and (d).

Baselines To illustrate the superiority of our approach, the following state-of-the-art methods are chosen as our baselines: 1) Strongly supervised fine-grained method Mask-CNN (Wei et al. 2018); 2) Weakly supervised fine-grained methods Bilinear CNN (Lin, RoyChowdhury, and Maji 2015), RA-CNN (Fu, Zheng, and Mei 2017), Filter-bank (Wang, Morariu, and Davis 2018), and Multi-attention (Zheng et al. 2017); 3) Semi-supervised fine-grained methods (Xu et al. 2016), (Niu, Veeraraghavan, and Sabharwal 2018), (Cui et al. 2016), and (Cui et al. 2018); 4) Web-supervised methods WSDG (Niu, Li, and Xu 2015), (Xiao et al. 2015), Decoupling (Malach and Shalev-Shwartz 2017), and Co-teaching (Han et al. 2018). For Co-teaching and Decoupling, we replace the basic network in them with the same backbone network VGG-16 as ours and train fine-grained models with the same web datasets. To be specific, we use the same implementation except for the batch size, which is changed to 64 and 16 in the first and second step, respectively. For Co-teaching, we set the maximum drop rate $\tau = 0.25$ and epoch $T_k = 10$. Experiments are conducted on two NVIDIA V100 GPU cards.

Experimental Results and Analysis

Table 1 presents fine-grained ACA results of various approaches on benchmark datasets. As demonstrated in Table 1, our proposed approach shows significant improvements, compared to other web-supervised methods on the CUB200 and Cars-196 datasets. On the FGVC-aircraft dataset, our approach achieves slightly better performance than Co-teaching.

Fig. 3 (a) presents the test accuracy vs. number of epochs of our approach, Decoupling, and Co-teaching on CUB200 dataset. From Fig. 3 (a), the memorization effect of networks can be clearly observed in our approach, whose test accuracy reaches a high level with fast speed and then gradually decreases. In contrast, the test accuracies of Decoupling and Co-teaching rise slowly with obvious fluctuation, failing to reach a high level at the early stage of training. This is because our approach has a better sample selection ability, which makes it able to reach a higher peak in much fewer epochs. Fig. 3 (b) shows the test accuracy vs. the number of epochs on CUB200, FGVC-Aircrafts, and Cars-196. By observing Fig. 3 (b), we can notice the same trend on the other two datasets.

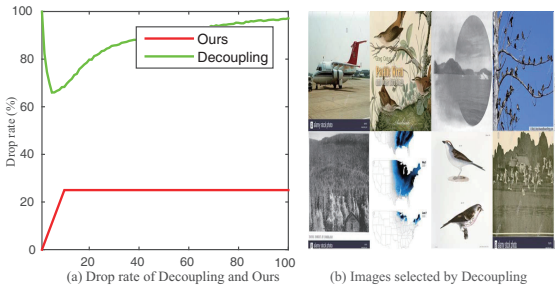

Figure 4: Drop rate of Decoupling (a) and sample images selected for training by Decoupling (b).

To further explain the sample selection ability of our approach, we record the selection result of each epoch during training since epoch $T > T_k$ (we set T_k=10) and compute the overlap rate of selected noisy samples. Given the number of overlapped noisy images N_o, training set $\mathcal{D}$ and drop rate $r(T)$, the overlap rate O is computed by $O = \frac{N_o}{|\mathcal{D}| \cdot r(T)}$. Fig. 3 (c) gives the result of our approach, Co-teaching, and Decoupling on CUB200 dataset. Fig. 3 (d) shows the result of our approach on three datasets.

Co-teaching leverages the same drop rate as ours. However, as the number of epochs increases, the overlap rate of all epochs in Co-teaching decreases to 0 rapidly, while our approach keeps a roughly stable number after a little drop (Fig. 3 (c)). Similarly, the overlap of contiguous epochs in Co-teaching keeps a small amount (around 10%), while our approach obviously holds more overlap, the number of which rises steadily as the number of epochs increases. That means our approach maintains a stable selection result and it becomes more stable as the training continues. This improvement benefits from our global selecting strategy. Specifically, Co-teaching performs sample selection in a mini-batch, where it can't tackle the noise rate imbalance problem. So its selection results are unstable and changing rapidly during training, further causes the network learning from the noisy images. By overcoming this drawback, our approach has

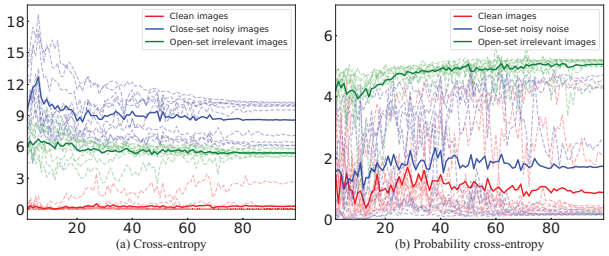

Figure 5: Cross-entropy (a) and Probability cross-entropy (b) of clean images, close-set noisy images and open-set irrelevant images. The value of each image is plotted in dotted line and the average value is plotted in solid line.

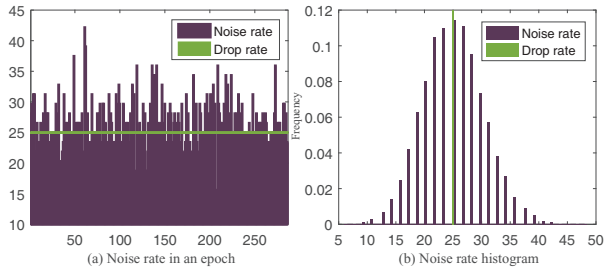

Figure 6: Noise rate of each mini-batch in an epoch (a) and noise rate histogram of all mini-batches(b).

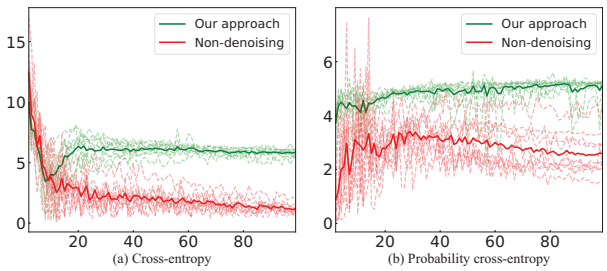

Figure 7: Cross-entropy (a) and Probability cross-entropy (b) of irrelevant noisy images in our approach and non-denoising training. The value of each image is plotted in dotted line and the average value is plotted in the solid line.

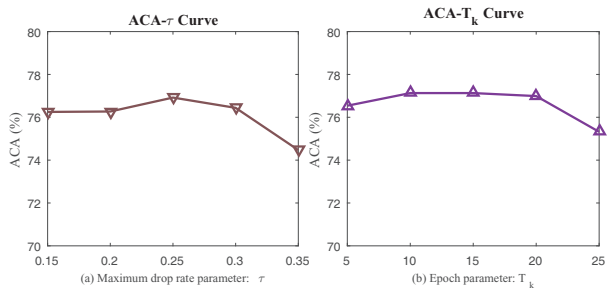

Figure 8: The parameter sensitivities of maximum drop rate τ and epoch T_k.

better sample selection consistency and better performance. From Fig. 3 (d), we can observe that our approach maintains stable sample selection results on all three datasets, especially on CUB200 and Cars-196.

Ablation Studies

Noise Rate Imbalance In this subsection, we investigate the noise rate imbalance problem in mini-batches with noisy bird training images. We record the number of dropped images during training in each mini-batch. Assuming that the dropped images are noisy samples, we can compute the noise rate R_i of mini-batch i. Given the number of dropped images N_i of mini-batch i and batch size N_b, we can calculate R_i by $R_i = \frac{N_i}{N_b}$. Fig. 6 (a) presents the noise rate of each mini-batch in a randomly selected epoch. It ranges from 12% to 42% with obvious fluctuation around the drop rate. To illustrate the distribution of noise rate R, we record noise rates of all mini-batches (more than 25000 mini-batches) during training and plot the histogram of R in Fig. 6 (b). By observing Fig. 6 (b), we can notice that R forms like a Gaussian distribution, ranging from 6% to 48%. Fig. 6 (a) explicitly shows the noise rate imbalance in mini-batches. Doing sample selection in a mini-batch with a fixed drop rate can't tackle the noise rate imbalance problem. While our proposed approach which leverages global sample selection can overcome it.

Decoupling selects samples with different predictions from two peer networks to update the network. It's a global selection method and doesn't have the noise rate imbalance problem. So it holds a stable selection result. As shown in Fig. 3 (c), the overlap rate of three contiguous epochs in Decoupling is close to our approach. However, the drop rate of Decoupling is too high. We record the number of dropped images and compute the drop rate of decoupling in each epoch. The result is shown in Fig. 4 (a). From Fig. 4 (a), we can observe that the drop rate of Decoupling is larger than 60% all the time and it climbs to nearly 100% as training continues.

The extreme high drop rate demonstrates that it can't make full use of the clean samples. Besides, since irrelevant noise is hard to fit, the peer networks have a high probability to produce different predictions of the irrelevant noisy samples. Then these noisy samples are used for training, misguiding the networks. Fig. 4 (b) visualizes some overlapped images which are used for training in Decoupling. These images are irrelevant noisy samples, indicating that Decoupling is not capable to tackle this irrelevant noise.

Probability Cross-entropy and Cross-entropy In this experiment, we compare the performance of probability cross-entropy and cross-entropy in identifying noise on noisy bird training images. We first save our models of each epoch during training. Then we leverage them to identify clean images, close-set noisy images, and open-set irrelevant images (30 images in total, 10 for each kind). We record their cross-entropy as well as probability cross-entropy. The experimental results are shown in Fig. 5. By observing Fig. 5 (b), we can notice that the probability cross-entropy of open-set irrelevant images is much larger than that of close-set noisy images and clean data. Compared with clean images, both close-set noisy images and open-set irrelevant images have a larger loss. From Fig. 5 (a) and (b), we can conclude

Table 2: ACA(%) performances of different backbones, dataset sizes, and frameworks.

Backbones	ACA (%)
VGG-16	77.22
VGG-19	75.87
ResNet-34	74.99

Dataset sizes	ACA (%)
50	71.87
75	74.85
100	77.22

Frameworks	ACA (%)
Co-teaching	75.46
Peer networks	76.30
Single network	77.22

that selecting samples with cross-entropy can't distinguish the noise of close-set and open-set. Nevertheless, leveraging our proposed probability cross-entropy to identify open-set irrelevant images is reliable.

Effectiveness of Denoising To explain the effectiveness of denoising proposed in our approach, we train the network using original web images without any denoising and record the probability cross-entropy and cross-entropy of 10 irrelevant noisy images during training. We also train the network using selected web images by our proposed approach and compare the results in Fig. 7. From Fig. 7 (b), we can find that the probability cross-entropy is large and declines extremely slowly during training, meaning that using probability cross-entropy can identify irrelevant noise during training and learn robust models. From Fig. 7 (a), we notice that the cross-entropy in non-denoising method gradually drops during training. In contrast, the cross-entropy in our approach drops slightly at first and then climbs to a roughly constant value. The explanation is that our approach has the ability to drop irrelevant noisy images before the network fit them.

Parameter sensitivity In parameter sensitivity analysis, we study the maximum drop rate τ and epoch T_k. As illustrated in Fig. 8, when τ increases from 0.15 to 0.3 or T_k rises from 5 to 20, the ACA performance remains roughly stable. It indicates that our approach is robust under real-world scenario. When τ is too large (τ=0.35), the ACA performance obviously drops. It may be caused by the fact that too many instances are dropped and network can not get sufficient training data. Similarly, too large T_k (T_k=25) degrades the performance. One possible explanation is that more irrelevant noisy images are learnt before they are dropped.

Influence of Different Backbones To investigate the influence of different CNN architectures, we replace VGG16 with VGG19 and ResNet34 (He et al. 2016). As shown in Table 2, these three backbone networks achieve similar performance on CUB200. The performances of VGG19 and ResNet34 are slightly worse than that of VGG16. One possible explanation is that we use the same coefficients setting for these backbones, which is best for VGG16.

Influence of Different Dataset Sizes We investigate the impact of data scale by changing the number of web images used for each category on CUB200. Specifically, we collect 50, 75, 100 images from the web for each category. As shown in Table 2, in general, the ACA performance improves steadily by using more training images. Therefore, web-supervised learning is a promising research direction as the large-scale dataset is easy to build through web images.

Influence of Multi-networks We conduct two experiments to study whether using multi-network can improve performance. In the first experiment, we combine our approach with Co-teaching framework, letting two networks select samples for each other. In the second experiment, we use two peer networks and leverage the outputs of them to compute probability cross-entropy. Given the softmax probability $\boldsymbol{p}(x_i)^{T-1}$ and $\boldsymbol{q}(x_i)^{T-1}$ of two peer networks, the cross-entropy $C(x_i)^T$ is computed by $C(x_i)^T = -\sum_{j=1}^{N} p_j(x_i)^{T-1} \log q_j(x_i)^{T-1}$. The results are demonstrated in Table 2. Both frameworks show worse performance than our proposed approach which only utilizes a single network. Compared with methods which need two networks, our approach is lighter and more efficient.

(a) Bird dataset (b) Aircraft dataset (c) Car dataset

Figure 9: Sample selection results on three web datasets. For each dataset, the irrelevant noisy and useful images selected by our approach are shown on the left and right, respectively.

Visualization Fig. 9 visualizes the sample selection results of our approach on three web datasets. From Fig. 9, we can observe that the irrelevant noisy images and useful images are clearly separated. Most irrelevant noisy images in the bird dataset have no relationship with birds. In the aircraft dataset, irrelevant noisy images are structure charts and cockpits, while in the car dataset, they tend to be logos and internal views of the car. They are related to the aircraft and car but different from images in standard dataset and harmful for training. Although irrelevant noisy images in these datasets are totally different, our approach is still able to distinguish them from different datasets. This selection results also demonstrate that our approach is robust and can be leveraged to refurbish the web images for practical applications.

Conclusion

In this paper, we presented a simple yet effective training method for web-supervised fine-grained classification tasks. Our key idea is to select samples that have a large probability cross-entropy as irrelevant noisy images and then drop them during training. Experiments on three real-world scenario datasets demonstrate that our approach has achieved the state-of-the-art performance.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 61976116, 61702163, 61702265, 61932020), National Key R&D Program of China (No. 2018AAA0102001), Sichuan Science and Technology Program (No. 2019YFG0003, 2018GZDZX0032), Natural Science Foundation of Jiangsu Province (No. BK20170856), and China Postdoctoral Science Foundation (No. 2019M651698).

References

Arpit, D.; Jastrzkebski, S.; Ballas, N.; Krueger, D.; Bengio, E.; Kanwal, M. S.; Maharaj, T.; Fischer, A.; Courville, A.; Bengio, Y.; et al. 2017. A closer look at memorization in deep networks. In *IMCL*, 233–242.

Bai, Y.; Zhang, Y.; Ding, M.; and Ghanem, B. 2018a. Finding tiny faces in the wild with generative adversarial network. In *CVPR*, 21–30.

Bai, Y.; Zhang, Y.; Ding, M.; and Ghanem, B. 2018b. Sod-mtgan: Small object detection via multi-task generative adversarial network. In *ECCV*, 206–221.

Cui, Y.; Zhou, F.; Lin, Y.; and Belongie, S. 2016. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. In *CVPR*, 1153–1162.

Cui, Y.; Song, Y.; Sun, C.; Howard, A.; and Belongie, S. 2018. Large scale fine-grained categorization and domain-specific transfer learning. In *CVPR*, 4109–4118.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Fu, J.; Zheng, H.; and Mei, T. 2017. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *CVPR*, 4438–4446.

Goldberger, J., and Ben-Reuven, E. 2016. Training deep neural-networks using a noise adaptation layer.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 8527–8537.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv*.

Huang, S.; Xu, Z.; Tao, D.; and Zhang, Y. 2016. Part-stacked cnn for fine-grained visual categorization. In *CVPR*, 1173–1182.

Jiang, L.; Zhou, Z.; Leung, T.; Li, L.-J.; and Fei-Fei, L. 2017. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. *arXiv*.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *ICCV*, 554–561.

Krause, J.; Sapp, B.; Howard, A.; Zhou, H.; Toshev, A.; Duerig, T.; Philbin, J.; and Fei-Fei, L. 2016. The unreasonable effectiveness of noisy data for fine-grained recognition. In *ECCV*, 301–320.

Liang, S.; Li, Y.; and Srikant, R. 2017. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv*.

Lin, D.; Shen, X.; Lu, C.; and Jia, J. 2015. Deep lac: Deep localization, alignment and classification for fine-grained recognition. In *CVPR*, 1666–1674.

Lin, T.-Y.; RoyChowdhury, A.; and Maji, S. 2015. Bilinear cnn models for fine-grained visual recognition. In *ICCV*, 1449–1457.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv*.

Malach, E., and Shalev-Shwartz, S. 2017. Decoupling" when to update" from" how to update". In *NeurIPS*, 960–970.

Niu, L.; Li, W.; and Xu, D. 2015. Visual recognition by learning from web data: A weakly supervised domain generalization approach. In *CVPR*, 2774–2783.

Niu, L.; Veeraraghavan, A.; and Sabharwal, A. 2018. Webly supervised learning meets zero-shot learning: A hybrid approach for fine-grained classification. In *CVPR*, 7171–7180.

Reed, S.; Lee, H.; Anguelov, D.; Szegedy, C.; Erhan, D.; and Rabinovich, A. 2014. Training deep neural networks on noisy labels with bootstrapping. *arXiv*.

Schroff, F.; Criminisi, A.; and Zisserman, A. 2011. Harvesting image databases from the web. *TPAMI* 33(4):754–766.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv*.

Song, H.; Kim, M.; and Lee, J.-G. 2019. Selfie: Refurbishing unclean samples for robust deep learning. In *ICML*, 5907–5915.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Wang, Y.; Liu, W.; Ma, X.; Bailey, J.; Zha, H.; Song, L.; and Xia, S.-T. 2018. Iterative learning with open-set noisy labels. In *CVPR*, 8688–8696.

Wang, Y.; Morariu, V. I.; and Davis, L. S. 2018. Learning a discriminative filter bank within a cnn for fine-grained recognition. In *CVPR*, 4148–4157.

Wei, X.-S.; Xie, C.-W.; Wu, J.; and Shen, C. 2018. Mask-cnn: Localizing parts and selecting descriptors for fine-grained bird species categorization. *PR* 76:704–714.

Xiao, T.; Xia, T.; Yang, Y.; Huang, C.; and Wang, X. 2015. Learning from massive noisy labeled data for image classification. In *CVPR*, 2691–2699.

Xu, Z.; Huang, S.; Zhang, Y.; and Tao, D. 2015. Augmenting strong supervision using web data for fine-grained categorization. In *ICCV*, 2524–2532.

Xu, Z.; Huang, S.; Zhang, Y.; and Tao, D. 2016. Webly-supervised fine-grained visual categorization via deep domain adaptation. *TPAMI* 40(5):1100–1113.

Yao, Y.; Zhang, J.; Shen, F.; Hua, X.; Xu, J.; and Tang, Z. 2017. Exploiting web images for dataset construction: A domain robust approach. *TMM* 19(8):1771–1784.

Yao, Y.; Shen, F.; Zhang, J.; Liu, L.; Tang, Z.; and Shao, L. 2018. Extracting privileged information for enhancing classifier learning. *TIP* 28(1):436–450.

Yao, Y.; Zhang, J.; Shen, F.; Liu, L.; Zhu, F.; Zhang, D.; and Shen, H. T. 2019. Towards automatic construction of diverse, high-quality image datasets. *TKDE*.

Zhang, N.; Donahue, J.; Girshick, R.; and Darrell, T. 2014. Part-based r-cnns for fine-grained category detection. In *ECCV*, 834–849.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2016. Understanding deep learning requires rethinking generalization. *arXiv*.

Zheng, H.; Fu, J.; Mei, T.; and Luo, J. 2017. Learning multi-attention convolutional neural network for fine-grained image recognition. In *ICCV*, 5209–5217.

Zhou, B.; Khosla, A.; Lapedriza, A.; Torralba, A.; and Oliva, A. 2016. Places: An image database for deep scene understanding. *arXiv*.