Supplementary Materials: Understanding and Utilizing Deep Neural Networks Trained with Noisy Labels

A. Further details on experiments

A.1. CIFAR-10

CIFAR-10 (Krizhevsky & Hinton, 2009) contains human-annotated labels which can be treated as true labels. To conduct experiments on the synthetic noisy labels, we randomly corrupt the labels according to a noise transition matrix T.

In all experiments, we set the batch size to 128, and implement (i) l_2 weight decay of 10^{-4} and (ii) data augmentation of horizontal random flipping and 32×32 random cropping after padding 4 pixels around images. In Sec. 5.1, we aim to verify our theory by demonstrating the worst case, so we use the ResNet-110 (He et al., 2016b) to ensure the model has the sufficiently high capacity to memorize all corrupted samples. While in Sec. 5.2 & 5.3, we use the ResNet-32 (He et al., 2016a) for the consideration of training efficiency.

In Sec. 5.2, we apply the Iterative Noisy Cross-Validation (INCV, Alg. 2) to select clean samples. For efficiency, we set the number of iterations to 4, and train the ResNet32 for 50 epochs at each iteration. We use the Adam optimizer with an initial learning rate 10^{-3}, which is divided by 2 after 20 and 30 epochs, and finally takes the value 10^{-4} after 40 epochs. In all other experiments, we train the networks for 200 epochs till convergence, using the Adam optimizer (Kinga & Adam, 2015) with an initial learning rate 10^{-3}, which is divided by 10 after 80, 120 and 160 epochs, and further divided by 2 after 180 epochs.

After selecting clean samples, we train DNNs robustly using Alg. 3. We set the warm-up epochs E_0 to 40 or 80 (i.e., 20% or 40% of the total number of training epochs) without fine tuning. If the size of the candidate set $\mathcal{C}$ is large, considering it has much more noisy labels than the selected relatively clean set $\mathcal{S}$, we set $E_0 = 80$ so that the network will focus on $\mathcal{S}$ until 80 epochs. Otherwise, we take $E_0 = 40$. In the INCV, we denote the proportion between the number of removed samples and selected samples as remove ratio r, which determines how many samples will be removed. We found that our algorithm is robust to r, which means slightly changing it does not affect the performance much. If we do not want to remove any samples, we set $r = 0$, otherwise we set $r = \frac{\varepsilon}{1-\varepsilon}$ without fine tuning, where ε is the estimated noise ratio of the original training set given by Alg. 2, hence $\frac{\varepsilon}{1-\varepsilon}$ is the proportion between the number of corrupted samples and the number of clean samples in the original training set.

For those baseline methods, there are many specific hyper-parameters, and we set the value according to their original papers. We train the same ResNet-32 for 200 epochs using the Adam optimizer with the same learning rate scheduler.

A.2. WebVision

To verify the practical usage of our method on real-world noisy labels, we use the WebVision dataset 1.0 (Li et al., 2017) which contains 2.4 million images crawled from the websites using the 1,000 concepts in ImageNet ILSVRC12 (Deng et al., 2009). The training set of the WebVision contains many real-world noisy labels. Since the dataset is quite large, for quick experiments, we use the first 50 classes of the Google image subset. We test the trained DNNs on the human-annotated WebVision validation set and the ILSVRC12 validation set.

We use the inception-resnet v2 (Szegedy et al., 2017). Following the standard training pipeline (Li et al., 2017), we first resize each image to make shorter size as 256. Then we implement standard data augmentation: randomly crop a patch of size 227×227 form each image, and horizontal random flipping is applied before feeding the patch to the network for training. The batch size is set to 128 for all experiments. We train the networks for 120 epochs using the SGD optimizer with an initial learning rate 0.1, which is divided by 10 after 40, and 80 epochs.

In our method, we first run the INCV to select clean samples. In the INCV, we set the number of iterations to 2, and train the model for simply 50 epochs at each iteration. We use the SGD optimizer with an initial learning rate 0.1, which is divided by 2 after 20 and 30 epochs, and finally takes the value 0.01 after 40 epochs. We set the remove ratio r to 0.1. After selecting clean samples, we train a model robustly using Alg. 3, where we set the warm-up epoch as $E_0 = 20$.

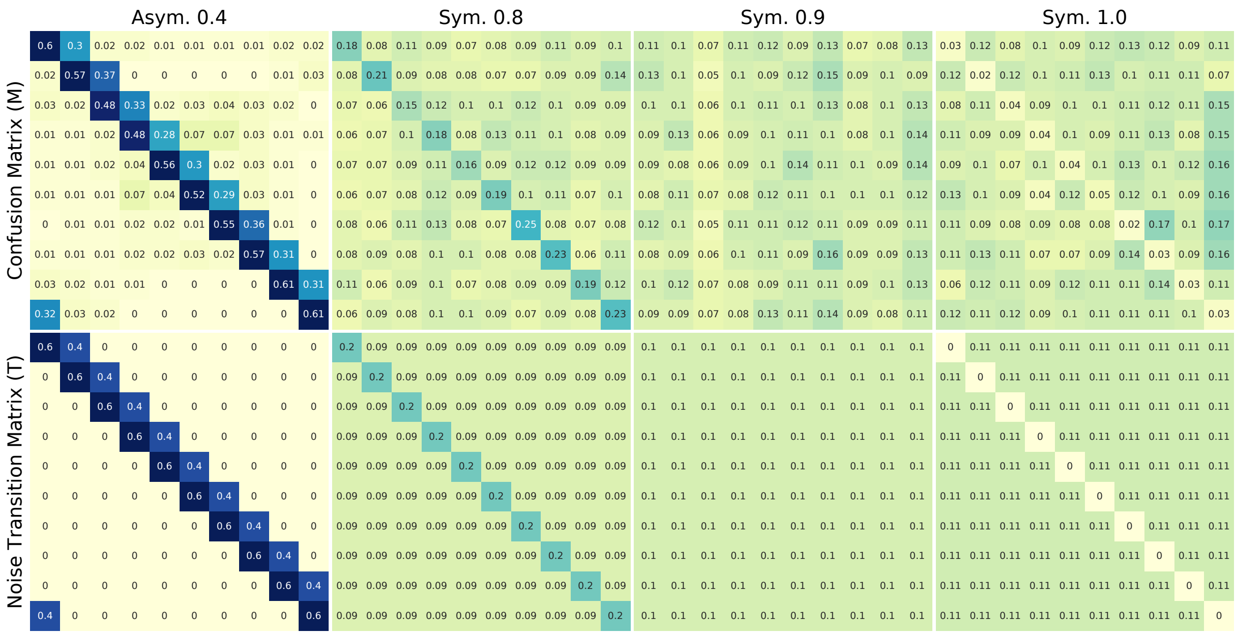

Figure 1. Confusion matrix (the first row) of ResNet-110 normally trained on corrupted CIFAR-10 with noise transition matrix T (the second row). We specifically examine the noise settings with low training accuracy. $M \approx T$ satisfies the statement presented in Claim 1.

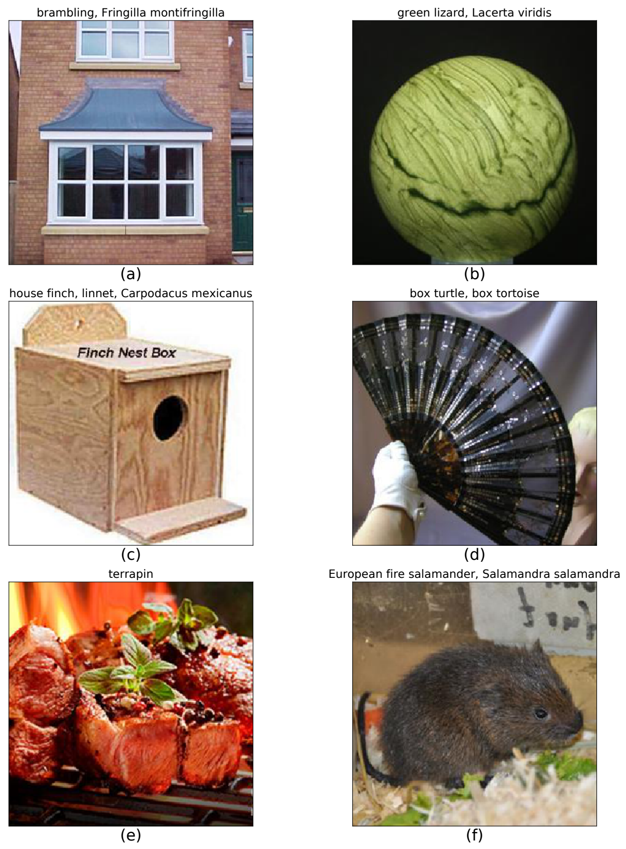
brambling, Fringilla montifringilla
(a)
green lizard, Lacerta viridis
(b)
house finch, linnet, Carpodacus mexicanus
Finch Nest Box
(c)
box turtle, box tortoise
(d)
terrapin
(e)
European fire salamander, Salamandra salamandra
(f)

Figure 2. Examples of automatically identified noisy labels in the WebVision dataset using the INCV. We annotate the labeled concepts on top of each image. The labels are obviously unreasonable.

B. More plots of the confusion matrix

We have shown in the main paper that when a network is trained with noisy labels, its confusion matrix M on the test set equals to the noise transition matrix T. This directly verifies our statement presented in Claim 1, which implies that the DNNs are able to fit the noisy training set exactly and generalize in distribution. Due to lack of space, in the main paper, we simply show results for symmetric noise of ratio 0.7. Here in Fig. 1, we show that $M \approx T$ holds for different noise types and noise ratios. We present the results for asymmetric noise of ratio 0.4, and then specifically investigate the noise settings which result in a low training accuracy, i.e., symmetric noise of ratio 0.8, 0.9 and 1.0 where the training accuracies are 0.40, 0.24 and 0.36. In this way, we also verify that the training accuracy converging to a extremely low value does not contradict our formulations on the generalization performance of DNNs trained with noisy labels.

C. The INCV automatically identifies many noisy labels in the WebVision dataset

In Sec. 5.3, we have demonstrated that on the WebVision dataset, compared with state-of-the-art methods, our training strategy is capable of training a model that achieves the best generalization performance on the clean validation set. In the experiments, we firstly select most clean samples out of the original training set use the Iterative Noisy Cross-Validation (INCV, Alg. 2). The INCV also identifies samples that are very likely to have a wrong label. In this Section, we demonstrate that the INCV does identify many noisy labels in the WebVision, as shown in Fig. 2. Since the images have different size with shorter size as 256, we crop each image from the center to form a square image. We first convert the observed label of each example to the correspond concept in the synsets, then annotate the concept on top of each image. In the WebVision, the 6 images are labeled as (a) brambling, Fringilla montifringilla; (b) green lizard, Lacerta viridis; (c) house finch, linnet, Carpodacus mexicanus; (d) box turtle, box tortoise; (e) terrapin; (f) European fire salamander, Salamandra salamandra; which are obviously unreasonable.

D. More discussions on Corollary 2.2

Without loss of generality, we assume $\forall i$, T_{ii} being the largest among $T_{ij}, j \in [c] := \{1, \cdots, c\}$. Based on Corollary 2.2 presented in the main paper, we can prove that under the cases of symmetric and asymmetric noise, Alg. 1 always selects a subset with smaller noise ratio than the original dataset, i.e., $\varepsilon_S < \varepsilon$, where ε is the noise ratio of the original dataset $\mathcal{D}$, and ε_S is the noise ratio of the selected set $\mathcal{S}$. Recall that $\varepsilon_S = 1 - LP$ according to the definition of LP.

For the symmetric noise, we have the definition $\forall i \in [c]$, $T_{ii} = 1 - \varepsilon$, and $T_{ij} = \varepsilon/(c-1), \forall j \neq i$. In this case, T_{ii} being the largest number among T_{ij} implies $\varepsilon/(c-1) < 1 - \varepsilon$. Using Eq. (10) in Corollary 2.2, we have

$$\begin{aligned} 1 - \varepsilon_S = LP &= \frac{(1-\varepsilon)^2}{(1-\varepsilon)^2 + \varepsilon^2/(c-1)} \\ &> \frac{(1-\varepsilon)^2}{(1-\varepsilon)^2 + \varepsilon(1-\varepsilon)} \\ &= 1 - \varepsilon. \end{aligned}$$

For the asymmetric noise, we have the definition $\forall i \in [c]$, $T_{ii} = 1 - \varepsilon$, $T_{ij} = \varepsilon$ for some $j \neq i$, and $T_{ij} = 0$ otherwise. In this case, T_{ii} being the largest number among T_{ij} implies $\varepsilon < 1 - \varepsilon$. Using Eq. (11) in Corollary 2.2, we have

$$\begin{aligned} 1 - \varepsilon_S = LP &= \frac{(1-\varepsilon)^2}{(1-\varepsilon)^2 + \varepsilon^2} \\ &> \frac{(1-\varepsilon)^2}{(1-\varepsilon)^2 + \varepsilon(1-\varepsilon)} \\ &= 1 - \varepsilon. \end{aligned}$$

Thus, we can conclude that $\varepsilon_S < \varepsilon$.

References

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. 2009.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *CVPR*, 2016a.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. pp. 630–645, 2016b.

Kinga, D. and Adam, J. B. A method for stochastic optimization. *ICLR*, 2015.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

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

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. Inception-v4, inception-resnet and the impact of residual connections on learning. 4:12, 2017.