Generation and Analysis of Feature-Dependent Pseudo Noise for Training Deep Neural Networks

Sree Ram Kamabattula¹ Kumudha Musini¹ Babak Namazi² Ganesh Sankaranarayanan² Venkat Devarajan¹.

Abstract-Training Deep neural networks (DNNs) on noisy labeled datasets is a challenging problem, because learning on mislabeled examples deteriorates the performance of the network. As the ground truth availability is limited with real-world noisy datasets, previous papers created synthetic noisy datasets by randomly modifying the labels of training examples of clean datasets. However, no final conclusions can be derived by just using this random noise, since it excludes feature-dependent noise. Thus, it is imperative to generate feature-dependent noisy datasets that additionally provide ground truth. Therefore, we propose an intuitive approach to creating feature-dependent noisy datasets by utilizing the training predictions of DNNs on clean datasets that also retain true label information. We refer to these datasets as "Pseudo Noisy datasets". We conduct several experiments to establish that Pseudo noisy datasets resemble feature-dependent noisy datasets across different conditions. We further randomly generate synthetic noisy datasets with the same noise distribution as that of Pseudo noise (referred as "Randomized Noise") to empirically show that i) learning is easier with feature-dependent label noise compared to random noise, ii) irrespective of noise distribution, Pseudo noisy datasets mimic feature-dependent label noise and iii) current training methods are not generalizable to feature-dependent label noise. Therefore, we believe that Pseudo noisy datasets will be quite helpful to study and develop robust training methods.

I. INTRODUCTION

Deep neural networks (DNNs) have demonstrated excellent performance in several tasks such as image classification [1], object detection [2] and several others, owing to the increasing size of training datasets [3] and advanced architectures [4]. However, the labeling process of the large datasets results in a large number of mislabeled examples in several application fields such as medical imaging [5], generative networks [6], etc. As a result, the generalization error increases, as the DNN learns on the mislabeled training examples [7]. Therefore, reducing the generalization error while training DNNs with noisy labeled datasets has become prominent work [8], [9]. Mislabeled examples are referred to as noise or label noise throughout this paper.

It is common practice to understand the learning behavior of clean and noisy examples using ground truth, i.e., true labels of training examples [10]–[13]. But unfortunately, most of the real-world noisy datasets do not provide ground truth [14]. So, previous papers created synthetic noisy datasets by *randomly* changing the true labels of some training examples in clean datasets with distributions such as symmetric and asymmetric while preserving ground truth [15]–[17].

Several observations and training methods have emerged by exploiting these synthetic noisy datasets. However, [18] pointed out a few contradictory findings and suggested that these methods might not be generalizable to realistic noisy datasets. For example, one such finding is that the generalization error increases, as the DNNs fit the noisy examples [7]. On the contrary, [19] shows that the DNNs can perform well even in the presence of label noise, when a slightly different noise distribution is adopted. Similarly, another popular small-loss observation becomes less effective with varying noise distributions [20].

On the other hand, based on these contradictory findings, [14] suggests that the label noise problem should be treated as three different sub problems, as each type has different characteristics. Specifically, in the order of increasing complexity, they are symmetric (random mislabeling), asymmetric (class-dependent) and feature-dependent (realistic) label noises.

Most of the current research is limited to symmetric and asymmetric noisy datasets due to the lack of ground truth with real world noisy datasets. Consequently, it is imperative to generate feature-dependent noisy datasets that additionally provide ground truth to study and develop robust training methods [21].

A. Analyzing learning behavior of label noise

DNNs learn easier patterns first and fit harder patterns in the later stages of training [22]. We utilize this fact to distinguish among the three noise categories (symmetric, asymmetric and feature-dependent noise) and, more importantly to understand the learning behavior of featuredependent label noise.

The noise is uniformly distributed among all the classes in symmetric noise, resembling the case of labelers randomly annotating the dataset [23]. Thus, the network needs to learn harder patterns to fit the *randomly* mislabeled examples. Since, the network fits on harder patterns in the later stages of training, easier patterns mostly refer to clean examples. Consequently, distinguishing between the clean and noisy examples with learning behavior of DNNs is easier with symmetric noise.

Distinguishing between the clean and noisy examples becomes harder with asymmetric noise compared to symmetric

Correspondence to sreeram.kamabattula@mavs.uta.edu ¹University of Texas at Arlington, Electrical/ Biomedical Engineering, Arlington, Texas, USA.

²UT Southwestern Medical Center, Dept. of Surgery, Dallas, Texas, USA. © 2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.



Fig. 1. Label recall (left), Accuracy (right) for symmetric on CIFAR-10 (top), Red-Mini-ImageNet (bottom) for $\tau = 0.2$ with ResNet32.

noise, as the noise is randomly distributed with just one class in an attempt to mimic the structure of realistic label noise. This can be related to real world noise caused by inexperienced labelers annotating the dataset, where only some randomness is introduced. For example, Red-MiniImageNet [18] real world noisy dataset belongs to this category. The authors have collected 161105 images by entering the 100 class names of Mini-ImageNet dataset in a Google search, and verified the labels by multiple human annotators. The verification process showed that there are 54,400 images with incorrect labels. Now, to create a $\tau\%$ noisy training dataset, $\tau\%$ of the original training images in the Mini-ImageNet dataset are *randomly* replaced by web images with incorrect labels. For testing, 5000 images in the ILSVRC12 validation set are used.

We plot the label recall (left) and accuracy (right) of symmetric noise and Red-MiniImageNet noisy dataset in Fig. 1. The experimental details are provided in section IV. With symmetric noise (top plots), it can be observed in the top left plot that the network learns on clean examples LR_{clean} (blue line) at a higher rate initially, while learning on noisy examples LR_{noisy} (orange line) remains minimal, obtaining the maximum test accuracy *MOTA* (vertical green line) in the initial stages of training. As the LR_{noisy} increases in the later stages of training, the test accuracy in the top right plot (orange line) keeps dropping after MOTA.

For Red-MiniImageNet noisy dataset, it can be observed in the bottom left plot that the network learns on both clean and noisy examples from the beginning of the training, unlike the symmetric noise. However, the rate of learning on noisy examples (orange line) is much lower compared to clean examples. As a result, the test accuracy keeps dropping after MOTA, as the train accuracy increases.

On the other hand, we noticed in Fig. 2, a different test accuracy behavior with ANIMAL-10N [20] real world noisy dataset, where the mislabeling happens only between confusing classes. It can be observed that both train and test



Fig. 2. Accuracy on ANIMAL-10N with CNN9.

accuracy are nearly constant in the later stages of training. So, the MOTA doesn't have much significance, since the test accuracy at the end of the training is also the same as that of test accuracy at MOTA. Unfortunately, the learning behavior of clean and noisy examples cannot be monitored, as the ground truth is not provided. However, based on the learning behavior observed in Fig. 1, we believe that this test accuracy behavior results when the network simultaneously learns on clean and noisy examples at a higher rate from the beginning of the training.

With feature-dependent label noise, as the mislabeled examples occur due to similarities between the features, *the network can easily fit both clean and noisy examples from the beginning of training*. In other words, easier patterns refer to both clean and noisy examples with feature-dependent label noise. Thus, distinguishing between clean and noisy examples based on learning behavior of DNNs is much harder. This analysis suggests the hypothesis that the network should easily fit the training examples with feature-dependent label noise compared to random noise.

B. Overview of our work

In this work, we establish a new class of noisy datasets called *Pseudo* noisy datasets. We refer to these as Pseudo noisy datasets because, we utilize the predictions of a DNN on the training examples of any clean dataset to create noisy datasets of desired noise rates. Thus, Pseudo noisy datasets can render true label information similar to the synthetic noisy datasets. However, it is important to emphasize that, in synthetic noisy datasets, a noise distribution is first chosen to randomly generate noisy labels. On the other hand, with Pseudo noisy datasets, noisy labels in fact arise due to an underlying learning misconception of the DNN. Thus, we claim that the Pseudo noisy datasets would be closer to feature-dependent label noise. We conduct several experiments to show that our created Pseudo noisy datasets resemble feature-dependent noisy datasets.

We further randomly generate synthetic noisy datasets by obtaining the noise distributions from Pseudo noisy datasets to fairly compare random and feature-dependent label noise. For convenience, we call the synthetic noisy datasets with the same noise distribution as that of Pseudo noisy datasets as *Randomized* noisy datasets.

We employ the learning behavior of DNNs on Randomized and Pseudo noisy datasets to i) prove our hypothesis that learning is easier with feature-dependent label noise compared to random noise and ii) prove that Pseudo noisy datasets imitate the feature-dependent label noise irrespective of the noise distribution.

Finally, we empirically show that the existing noise-robust training methods are not generalizable to feature-dependent label noise. Thus, we believe that Pseudo noisy datasets will be quite helpful in developing effective training methods.

II. RELATED WORK

Our paper focuses on label noise, especially Pseudo noise. To the best of our knowledge, pseudo labels have been used for training in the past [24]–[26], but they have never been utilized to create feature-dependent noisy datasets.

Web label noise and synthetic noise are closely related to our paper. In web noisy datasets, labeling is often performed by crowd sourcing [27], online queries [28], etc. For example, several clothing images are gathered from online shopping websites and labels are automatically assigned from surrounding texts in Clothing1M noisy dataset [9]. A few other widely used web noisy datasets are ANIMAL-10N [20], Food101N [29] and WebVision [30]. However, these noisy datasets either do not provide ground truth or only a small clean validation set is available. Therefore, synthetic noisy datasets are exploited to develop different training methods.

One common approach is to select clean samples and train the network on the selected samples [31]–[33]. A few methods correct the loss function based on the noise estimation [16] and [34], and assigning higher weights to clean samples [35] and [29]. Some research focuses on i) developing noise-robust loss functions [17], [36] and [37], ii) identifying an early training stop point [10], [13] and [38] and iii) semi-supervised learning [18] and [39]. However, the conclusions cannot be derived using these synthetic noisy datasets as they are just random noise [18]. Therefore, [14], [40] and [41] developed label-corruption algorithms to synthetically create feature dependent noisy datasets.

For example, the authors of [14] synthetically generated feature-dependent noise, where the probability of mislabeling depends on the similarities in features of training examples. However, it can be observed in the left plot of Fig. 3 that the learning on noisy examples (orange line left plot) increases in the later stages of training, not satisfying the previously discussed learning behavior of feature-dependent label noise. Therefore, we create feature-dependent noisy datasets where the network significantly learns on both clean and noisy examples from the beginning of training. Our proposed approach also allows other researchers to easily create their own feature-dependent noisy datasets.

III. PSEUDO AND RANDOMIZED NOISY DATASETS

A. Motivation of Pseudo noise

The key idea of developing Pseudo noisy datasets is inspired from the following human analogy. A person annotating a given dataset is likely to label an example correctly or incorrectly based on his proficiency (expert, trainee, etc.). For example, i) the labeler might be unable to distinguish among examples of different classes with similar features, ii) the labeler might find a specific class difficult to classify, etc. So, labelers with diverse expertise will annotate the same dataset with different percentages of noise rate and distributions.

Now, we relate the labeler to a DNN. The labels of a dataset annotated by a person with no prior knowledge is analogous to the predictions of a DNN with initialized weights. Furthermore, the DNN's weights at subsequent epochs can be comparable to expertise of different labelers. Therefore, we believe that the prior knowledge of a labeler can be equivalent to the learning distribution of the DNN on a training dataset (represented by the weights of the DNN). In other words, a labeler will annotate examples of a given dataset correctly or incorrectly based on their prior knowledge. Similarly, the expected prediction of the DNN on a new dataset will follow the distribution of the learning on the training dataset.

Based on this analysis, we can utilize the prediction errors of the DNNs on the training dataset to create noisy datasets. Thus, we train a DNN on a dataset with known ground truth and store the predictions of the network for each epoch. We further select the predictions at specific percentages of training accuracy to create Pseudo noisy datasets of desired noise rates. Therefore, the examples are mislabeled due to underlying learning misconception of DNN in Pseudo noisy datasets. Thus, we claim that the Pseudo noisy datasets would be closer to the feature-dependent label noise. Later, we will experimentally support this claim.

B. Pseudo noise generation

We select the predictions of the network on training examples at 1- τ % training accuracy to create a Pseudo noisy dataset with a noise rate of τ . However, it should be noted that, 1- τ % training accuracy can be obtained at multiple epochs resulting in several Pseudo noisy datasets with the same noise rate τ . Specifically, two factors vary in such noisy datasets with identical τ : the, i) distribution of clean examples and ii) distribution of training examples, for each class. Thus, we define two parameters α and β , which measure the distribution of clean examples and training examples over all the classes respectively, to create Pseudo noisy datasets with identical τ and distinct noise distributions, as shown in Table I. This can be related to creating different types of synthetic noisy datasets (symmetric and asymmetric) with identical τ .

Let \hat{Y} and Y represent the true labels and noisy labels of training examples respectively in a Pseudo noisy dataset. Let N_{ij} denote the noise distribution, where *i* represents the



Fig. 3. Label recall (left), Accuracy (right) for Feature-dependent noise on CIFAR-10 $\tau = 0.2$ with ResNet32.

TABLE I TWO NOISE DISTRIBUTIONS OF 0.2 τ .

Ŷ	Y					Ŷ		Y			
	Α	В	C	D	ĺ		А	В	C	D	
Α	0.86	0.04	0.06	0.04	ĺ	Α	0.81	0.11	0.04	0.04	
В	0.1	0.77	0.07	0.06		В	0.02	0.98	0	0	
C	0.08	0.13	0.68	0.11		C	0.32	0.09	0.53	0.06	
D	0	0.06	0.1	0.84		D	0.08	0.04	0	0.88	

TABLE II ALPHA AND BETA VALUES IN OUR EXPERIMENTS.

Dataset	CIFA	R-10	CIFAR-100		
	0.2	0.5	0.2	0.5	
α	0.15	0.21	0.12	0.22	
β	0.31	0.4	0.29	0.78	

true class and *j* denotes the corresponding class in Y. α is calculated by taking the standard deviation (σ) of N_{ij} at i = j, for all classes. Similarly, β is found by calculating the σ of distribution of training examples with reference to Y for each class, represented by N_j . In simple words, N_j is obtained by calculating the column-wise sum of N_{ij} .

$$\alpha = \sigma(diag(N_{ij})) \qquad \beta = \sigma(N_j) \qquad (1)$$

In this work, we conduct experiments with shown α and β values in Table II.

C. Randomized noise generation

To generate Randomized noisy datasets, examples are mislabeled by following a predetermined noise distribution obtained from Pseudo noisy datasets. Therefore, both Randomized and Pseudo noisy datasets have the exact same N_{ij} (either left or right Table I). However, the key difference is in the mislabeled examples. In Randomized noise, examples to be mislabeled are randomly picked, whereas in Pseudo noise, mislabeling of examples will occur due to underlying learning misconception of the DNN.

IV. EXPERIMENTS

We train the network with the Adam optimizer, momentum of 0.9, batch size of 128 for 200 epochs with ResNet32 architecture. The initial learning rate is set to 0.001 and multiplied by 0.5, 0.25, 0.1 at 20, 30 and 40 epochs respectively for all our experiments. We create several Pseudo noisy datasets of varied noise rates τ from two clean benchmark datasets: CIFAR-10 and CIFAR-100.

Metrics: Clean label recall at each epoch is calculated by dividing the total number of correctly predicted clean examples at that epoch over the total number of clean examples in the dataset. Likewise, noisy label recall is calculated. Let LR_{clean} and LR_{noisy} denote the vectors of clean and noisy label recall values respectively for all the epochs.

A. Evaluation of Pseudo noise

We plot the label recall and accuracy of Randomized (top) and Pseudo (bottom) noise in Fig.4. It can be observed in the top left plot that LR_{noisy} remains minimal in the initial stages of training similar to symmetric noise (Fig. 1). As the



Fig. 4. Label recall (left) and Accuracy (right) on CIFAR-10 with ResNet32 for $\tau = 0.2$: Symmetric (top), Randomized (middle) and Pseudo (bottom).



Fig. 5. Label recall (left) and Accuracy (right) on CIFAR-100 with ResNet32 for $\tau = 0.5$: Randomized (top) and Pseudo (bottom).

 LR_{noisy} increases in the later stages, the test accuracy keeps dropping as shown in the top right plot.

Pseudo noise on the other hand exhibits a different learning behavior as can be seen in the bottom plots. It can be observed that the network fits both clean (LR_{clean}) and noisy (LR_{noisy}) examples from the beginning of the training at a higher rate with Pseudo noise. Consequently, it can be noticed that the test accuracy (bottom right) does not vary much in the later stages of training. This clearly satisfies the previously discussed learning behavior of feature-dependent label noise.

It should be noted that the τ and noise distribution in the top and bottom plots are exactly the same. The only difference is that the examples are randomly mislabeled in Randomized noise, whereas examples are mislabeled due to underlying misconception of DNN in Pseudo noise.

The above discussed learning behavior of DNNs on Randomized and Pseudo noise is consistent across different conditions: i) higher τ and harder dataset (Fig. 5), ii) different



Fig. 6. Label recall (left) and Accuracy (right) on CIFAR-100 for $\tau = 0.5$ Pseudo: CNN9 (top) and ResNet110 (bottom).



Fig. 7. Label recall (left) and Accuracy (right) on CIFAR-10 for $\tau = 0.2$ Pseudo with ResNet32: constant learning rate (top) and decay 2 (bottom).

architectures (Fig. 6) and iii) varying learning rate (Fig. 7). The top plot of Fig. 7 is with constant initial learning rate (0.001), while the bottom plot is obtained when the initial learning rate is multiplied by 0.1 at 80, 120, 160 epochs and 0.5 at 180^{th} epoch (referred as decay 2 in figure). Note that the drop towards the end is due to the change in learning rate. These results clearly indicates that Pseudo noisy datasets resemble feature-dependent noisy datasets.

B. Validity of Pseudo noise

In the plots (Fig 4-5), we previously noticed that the learning behavior of Randomized noisy datasets doesn't resemble feature-dependent noise. This suggests that the noisy datasets created with *predetermined noise distributions* don't mimic the feature-dependent label noise, even when the noise distribution is same as Pseudo noise.

On the other hand, we verify the learning behavior of Pseudo noise when the noise distribution is similar to *symmetric* noise. It can be observed in Fig. 8 ($\alpha = 0.11$, $\beta = 0.15$) that the network still learns on both clean and



Fig. 8. Accuracy with ResNet32: CIFAR-10 for $\tau = 0.18$ Pseudo (left).

noisy examples from the beginning of training. This clearly shows that irrespective of the noise distribution, Pseudo noisy datasets are feature-dependent noisy datasets.

C. Learning is easier with feature-dependent label noise

We further utilize the learning behavior of DNNs to support our hypothesis that learning is easier with featuredependent label noise compared to Random noise.

It can be observed in the top plots of Fig.5 that the training accuracy (top right blue line) of Randomized noise for harder dataset CIFAR-100 and higher noise rate τ of 0.5 is very low, as the network needs to learn harder patterns to find relationship among the clean examples and randomly mislabeled examples. As a result, the network fits mostly clean examples in the initial stages.

On the other hand, the right bottom plot of Fig.5 clearly show that the training accuracy is higher for Pseudo noise despite harder dataset and higher noise rate. This indicates that the network can easily fit both clean and noisy examples by learning easier patterns from the beginning of training, as the mislabeling is caused due to underlying misconception of the DNN. Therefore, this proves that learning with Pseudo noisy datasets (feature-dependent label noise) is much easier compared to random noisy datasets.

D. Comparison of existing noise-robust training methods

In this section, we compare several existing training methods on Pseudo and Randomized noise with CIFAR-10 and CIFAR-100 on different τ values as shown in Table III. In particular, we compare the following broad approaches: i) sample selection methods (Co-teaching [15] and INCV [42]), ii) loss correction methods (F-correction [16], Bootsoft [25] and Joint [26]), iii) robust loss functions (SCE [17] and GCE [36]), iv) early stopping (AutoTSP [13] and NHA [38]) and v) standard training without early stopping (Standard).

We note the test accuracy and label recall at two points: MOTA (where maximum test accuracy is obtained) and final epoch (referred as MOTA/ final epoch in the Table) for all the approaches, except the early stopping methods (accuracy at stop point is reported).

As discussed earlier, the MOTA and the end test accuracy of Standard method drastically vary for Randomized noise, while the variance is small for Pseudo noise. It can also be observed that the MOTA obtained for Randomized noise is much higher than the Pseudo for CIFAR-10 dataset, because higher noisy label recall is obtained for Pseudo compared to Randomized noise as shown in Table IV. However, the

TABLE III

AVERAGE TEST ACCURACY (%, 3 RUNS) COMPARISON OF RANDOMIZED AND PSEUDO NOISE WITH RESNET32 (MAXIMUM/FINAL TEST ACCURACY).

Dataset		CIFA	R-10		CIFAR-100				
Noise rate	0.2		0.5		0.2		0.5		
Noise type	Randomized	Pseudo	Randomized	Pseudo	Randomized	Pseudo	Randomized	Pseudo	
Standard	84.37/76.75	78.85/75.17	70.13/56.74	52.49/50.99	57.63/52.13	58.71/56.75	43.4/36.26	45.23/44.35	
Co-teaching [15]	88.45/ 88.01	80.1/ 79.07	67.26/67.11	51.1/50.32	62.41/ 62.08	61.08/ 60.97	45.4/ 44.85	44/43.59	
INCV [42]	88.13/87.71	78.78/77.69	68.4/ 68.3	52.25/51.22	58.29/58.07	60.25/60	40.12/40	44.4/43.57	
F-correction [16]	84.72/76.7	78.35/74.67	68.82/54.85	53.62/51.06	56.95/53.02	58.73/56.76	43.72/35.39	44.6/42.9	
SCE [17]	83.78/82.32	78.22/73.79	66.02/64.77	52.14/50.99	54.6/53.41	55.15/54.53	40.37/38.44	42.67/41.92	
GCE [36]	-	79.57/74.81	-	52.56/50.23	-	57.04/55.12	42.26/40.6	44.55/41.38	
Joint [26]	86.7/80.96	79.01/74.01	78.56/68.11	53.43/51.37	57.33/54.25	58.59/57.73	49.13/44.18	46.77/43.96	
Bootsoft [25]	84.46/75.41	78.83/77.16	68.34/57.84	52.52/51.31	57.96/52.13	58.71/56.71	44/38.24	44.89/43.73	
AutoTSP [13]	81.21	75.51	63.32	52.49	55.73	56.44	43.29	44.06	
NHA [38]	81.07	74	64.61	42.24	52.13	56.3	39.33	35.57	

TABLE IV

NOISY LABEL RECALL OF RANDOMIZED AND PSEUDO NOISE AT MAXIMUM/ FINAL TEST ACCURACY WITH RESNET32.

Dataset		CIFA	R-10		CIFAR-100			
Noise rate	0.2		0.5		0.2		0.5	
Noise type	Randomized	Pseudo	Randomized	Pseudo	Randomized	Pseudo	Randomized	Pseudo
Standard	0.12/0.49	0.78/0.9	0.14/0.45	0.76/0.84	0.11/0.32	0.53/0.76	0.14/0.3	0.65/0.82
Co-teaching [15]	0.09/0.12	0.48/0.67	0.13/0.14	0.46/0.61	0.07/0.19	0.31/0.62	0.09/0.2	0.31/0.54
SCE [17]	0.09/0.16	0.65/0.74	0.13/0.19	0.71/0.75	0.07/0.13	0.34/0.51	0.13/0.22	0.54/0.68

maximum test accuracy is similar for both Pseudo and Randomized with CIFAR-100. It is likely because, generalization becomes harder for higher number of classes.

Among the existing training methods that we compared, it can be observed that the sample selection methods outperform all the others on both Pseudo and Randomized noise. Specifically, co-teaching consistently achieves higher test accuracy across different noise ratios and datasets, while INCV becomes inaccurate with higher number of classes. It can be further observed that the generalization performance is not improved than Standard with the remaining approaches for both Pseudo and Randomized noise. It is likely because, the methods are sensitive to small perturbations in their assumed noise distributions. Among these remaining methods, i) the test accuracy of SCE remains consistent from MOTA till the end of training across different cases, while it keeps dropping for other methods much like Standard training and ii) AutoTSP achieves higher accuracy at the end of training for 0.5 τ on CIFAR-10 and CIFAR-100 than the other approaches, with just finding a training stop point on Standard training, unlike the remaining approaches which modify the training framework.

As previously mentioned, co-teaching achieves higher test accuracy than the Standard even for Pseudo noise. However, the LR_{noisy} in Table IV implies that co-teaching still significantly learns on noisy examples with Pseudo noise, while it remains minimal with Randomized noise. This indicates that the current training methods are not generalizable to feature-dependent noisy datasets.

These results altogether substantiate the significance of Pseudo noise. We believe that our proposed Pseudo noisy datasets together with the ground truth information can be utilized to study and develop robust training methods with feature-dependent label noise.

V. CONCLUSION

In this work, we proposed Pseudo noisy datasets by utilizing the predictions of DNNs that also provide ground truth to effectively study feature-dependent label noise. We first established through the learning behavior of DNNs that Pseudo noisy datasets mimic feature-dependent label noise across different conditions, where distinguishing between clean and noisy examples is harder. We further generated a Randomized noise with the same noise distribution as that of Pseudo noise and empirically showed that i) learning is easier with feature-dependent label noise compared to random noise, ii) regardless of the noise distribution, pseudo noisy datasets resemble feature-dependent label noise and iii) several existing noise-robust training methods are not generalizable to feature-dependent (Pseudo) label noise. Therefore, we believe that our proposed approach will allow other researchers to create their own feature-dependent noisy datasets effortlessly in various domains in order to develop effective training methods.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in Advances in Neural Information Processing Systems, 2012. [Online]. Available: http://code.google.com/p/cuda-convnet/

- [2] J. Redmon, S. Divvala, R. B. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779– 788, 2016.
- [3] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference* on Computer Vision and Pattern Recognition. Institute of Electrical and Electronics Engineers (IEEE), 3 2010, pp. 248–255.
- [4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. [Online]. Available: http://image-net.org/challenges/LSVRC/2015/
- [5] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," *Medical Image Analysis*, vol. 65, p. 101759, 10 2020.
- [6] T. Kaneko, Y. Ushiku, and T. Harada, "Label-noise robust generative adversarial networks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019.
- [7] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," 5th International Conference on Learning Representations, ICLR 2017 -Conference Track Proceedings, 2017.
- [8] G. Algan and I. Ulusoy, "Image Classification with Deep Learning in the Presence of Noisy Labels: A Survey," Arxiv Preprint arXiv:1912.05170, 2019. [Online]. Available: http://arxiv.org/abs/ 1912.05170
- [9] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang, "Learning from Massive Noisy Labeled Data for Image Classification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2691–2699.
- [10] X. Ma, Y. Wang, M. E. Houle, S. Zhou, S. M. Erfani, S.-T. Xia, S. Wijewickrema, and J. Bailey, "Dimensionality-Driven Learning with Noisy Labels," in 35th International Conference on Machine Learning, ICML 2018, vol. 8. International Machine Learning Society (IMLS), 2018, pp. 5332–5341.
- [11] Y. Sun, Y. Tian, Y. Xu, and J. Li, "Limited Gradient Descent: Learning with Noisy Labels," *IEEE Access*, vol. 7, pp. 168 296–168 306, 2019.
- [12] G. Pleiss, T. Zhang, E. E. Asapp, and K. Q. Weinberger, "Identifying Mislabeled Data using the Area Under the Margin Ranking," in *Neural Information Processing Systems*, 2020.
- [13] S. R. Kamabattula, V. Devarajan, B. Namazi, and G. Sankaranarayanan, "Identifying Training Stop Point with Noisy Labeled Data," *Arxiv Preprint arXiv:2012.13435*, 12 2020. [Online]. Available: https://arxiv.org/abs/2012.13435
- [14] G. Algan and Ä. Ulusoy, "Label Noise Types and Their Effects on Deep Learning," arxiv 2003.10471, 3 2020. [Online]. Available: http://arxiv.org/abs/2003.10471
- [15] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. W. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Advances in Neural Information Processing Systems*, 2018.
- [16] G. Patrini, A. Rozza, A. K. Menon, R. Nock, and L. Qu, "Making Deep Neural Networks Robust to Label Noise: a Loss Correction Approach," in *The IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2017, pp. 1944–1952.
- [17] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey, "Symmetric cross entropy for robust learning with noisy labels," *arXiv*, pp. 322– 330, 2019.
- [18] L. Jiang, D. Huang, M. Liu, and W. Yang, "Beyond Synthetic Noise: Deep Learning on Controlled Noisy Labels," *37th International Conference on Machine Learning, ICML*, 2020. [Online]. Available: http://arxiv.org/abs/1911.09781
- [19] D. Rolnick, A. Veit, S. Belongie, and N. Shavit, "Deep Learning is Robust to Massive Label Noise," *Arxiv Preprint arXiv:1705.10694*, 2017.
- [20] H. Song, M. Kim, and J.-G. Lee, "SELFIE: Refurbishing Unclean Samples for Robust Deep Learning," in the 36th International Conference on Machine Learning (ICML), 2019.
- [21] H. Song, M. Kim, D. Park, and J.-G. Lee, "Learning from Noisy Labels with Deep Neural Networks: A Survey," arxiv, 7 2020. [Online]. Available: http://arxiv.org/abs/2007.08199
- [22] D. Arpit, S. Jastrzębski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, and S. Lacoste-

Julien, "A Closer Look at Memorization in Deep Networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, 2017.

- [23] A. Khetan, Z. C. Lipton, and A. Anandkumar, "Learning from noisy singly-labeled data," 2017.
- [24] Y. Ding, L. Wang, D. Fan, and B. Gong, "A Semi-Supervised Two-Stage Approach to Learning from Noisy Labels," *arxiv* 1802.02679, 2 2018. [Online]. Available: http://arxiv.org/abs/1802.02679
- [25] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich, "Training Deep Neural Networks on Noisy Labels with Bootstrapping," in 3rd International Conference on Learning Representations, ICLR 2015 - Workshop Track Proceedings. International Conference on Learning Representations, ICLR, 2015.
- [26] D. Tanaka, D. Ikami, T. Yamasaki, and K. Aizawa, "Joint Optimization Framework for Learning with Noisy Labels," in *The IEEE Conference* on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5552– 5560.
- [27] Y. Yan, R. Rosales, G. Fung, R. Subramanian, and J. Dy, "Learning from multiple annotators with varying expertise," *Machine Learning*, vol. 95, no. 3, pp. 291–327, 2014.
- [28] A. Blum, A. Kalai, and H. Wasserman, "Noise-tolerant learning, the parity problem, and the statistical query model," *Journal of the ACM*, vol. 50, no. 4, pp. 506–519, 2003.
- [29] K.-H. Lee, X. He, L. Zhang, L. Yang, and J. D. A. I. Research, "CleanNet: Transfer Learning for Scalable Image Classifier Training with Label Noise," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5447–5456.
- [30] W. Li, L. Wang, W. Li, E. Agustsson, and L. Van Gool, "WebVision Database: Visual Learning and Understanding from Web Data," arXiv: 1708.02862, 8 2017. [Online]. Available: http://arxiv.org/abs/1708.02862
- [31] E. Malach and S. Shalev-Shwartz, "Decoupling "when to update" from "how to update"," in Advances in Neural Information Processing Systems 30 (NIPS 2017), 2017, pp. 960–970.
- [32] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei, "MentorNet: Learning Data-Driven Curriculum for Very Deep Neural Networks on Corrupted Labels," in 35th International Conference on Machine Learning, ICML, vol. 5, 2018, pp. 3601–3620.
- [33] H.-S. Chang, E. Learned-Miller, and A. Mccallum, "Active Bias: Training More Accurate Neural Networks by Emphasizing High Variance Samples," in Advances in Neural Information Processing Systems 30 (NIPS 2017), 2017, pp. 1002–1012.
- [34] J. Goldberger and E. Ben-Reuven, "TRAINING DEEP NEURAL-NETWORKS USING A NOISE ADAPTATION LAYER," in International Conference on Learning Representations, 2017.
- [35] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to Reweight Examples for Robust Deep Learning," in 35th International Conference on Machine Learning, ICML 2018, vol. 10. International Machine Learning Society (IMLS), 2018.
- [36] Z. Zhang and M. R. Sabuncu, "Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels," in Advances in Neural Information Processing Systems 31 (NIPS 2018), 2018, pp. 8778–8788.
- [37] X. Wang, Y. Hua, E. Kodirov, and N. M. Robertson, "IMAE for noiserobust learning: Mean absolute error does not treat examples equally and gradient magnitude's variance matters," in *arXiv*, 2019.
- [38] H. Song, M. Kim, D. Park, and J.-G. Lee, "Prestopping: How Does Early Stopping Help Generalization against Label Noise?" arXiv:1911.08059, 2019. [Online]. Available: http://arxiv.org/abs/ 1911.08059
- [39] J. Li, R. Socher, and S. C. H. Hoi, "DIVIDEMIX: LEARNING WITH NOISY LABELS AS SEMI-SUPERVISED LEARNING," in International Conference on Learning Representations (ICLR), 2020.
- [40] X. Xia, T. Liu, B. Han, N. Wang, M. Gong, H. Liu, G. Niu, D. Tao, and M. Sugiyama, "Parts-dependent Label Noise: Towards Instancedependent Label Noise," in *Neural Information Processing Systems*, no. NeurIPS, 2020, pp. 1–14.
- [41] H. Cheng, Z. Zhu, X. Li, Y. Gong, X. Sun, and Y. Liu, "Learning with Instance-Dependent Label Noise: A Sample Sieve Approach," arXiv 2010.02347v1, 2020.
- [42] P. Chen, B. Liao, G. Chen, and S. Zhang, "Understanding and Utilizing Deep Neural Networks Trained with Noisy Labels," in 36th International Conference on Machine Learning, ICML. International Machine Learning Society (IMLS), 2019.