Final Report of Analyzing the Effect of Label Noise on Selective Classification for Deep Neural Networks

Jun Chen **Yitong Chen** **Yuhao Zhao** **Yifan Zheng**
University of Southern California
{*jxchen, chenyito, zhaoyuha, yzheng84*}*@usc.edu*

1 Abstract

Selective prediction provides deep learning models the ability to refrain from making a choice (Geifman and El-Yaniv, 2017). By using a Maximum Softmax Probability (MaxProb) to empirically select a confidence level, we can determine a risk-coverage graph on how the confidence level affects the model performance. However, noise is everywhere. Label noise occurs when a human inaccurately annotates a sample. For example, human can misclassify a cat as a dog. The implemented noises are uniform and asymmetric. The result shows that label noise does impact selective prediction. Selective prediction performs better on uniform noise over asymmetric noise.

2 Introduction

The objective of our experiments is to examine the effect of the label noise on the selective classification for deep neural networks (Geifman and El-Yaniv, 2017). The label noise refers to incorrect or inaccurate label, which is common in real datasets. The selective classification refers to 'classification with a reject option', in which situation there will always be a trade-off between coverage and accuracy. The risk-coverage curve is usually used to visualize this trade-off (El-Yaniv and Wiener, 2010a). We focus on two computer vision tasks, the image classification and the semantic segmentation. First, for the image classification, we reproduce and create the baseline performance of VGG-16 (Karen Simonyan, 2015) on MNIST (LeCun and Cortes, 2010), CIFAR-10, and CIFAR-100 datasets (Krizhevsky et al., 2009); for the semantic segmentation, we build two fully convolutional neural networks (FCN8 and FCN16) (Long et al., 2014) on the Kitti dataset (Alhaija et al., 2018). Then, we introduce different types and amount of label noise into the *training* data of each dataset. We use the performance evaluation metrics of selective classification to compare the models trained with and without label noise. Finally, we analyze the experiment results and decide what kind of improvement we can do.

3 Related Work

There are mainly 5 papers we are referring to. Their works give us the great inspiration on our experiments. First, it is "On the Foundations of Noise-free Selective Classification" (El-Yaniv and Wiener, 2010b), which talks about the essence in selective classification is the trade-off classifier coverage for higher accuracy. It gives us the idea about the use of selective classification and risk-coverage (RC) trade-off in our experiment. The second one is "Selective Classification for Deep Neural Networks" (Geifman and El-Yaniv, 2017). It mainly focus on how Selective classification techniques used in deep neural networks (DNNs), which helps us use SR/MaxProb as the baseline to measure the experiment. The third one is "Investigating Selective Prediction Approaches Across Several Tasks in IID, OOD, and Adversarial Settings" (Varshney et al., 2022), which focus on which selective prediction approaches work best across tasks. It tells us to do selective prediction(SP) evaluation experiments across tasks and settings to better evaluate the performance of the selective prediction. The fourth is "No True State-of-the-Art? OOD Detection Methods are Inconsistent across Datasets" (Fahim Tajwar, 2021), which claims that OOD Detection Methods are Inconsistent across Different Datasets. It suggests us to do selective prediction evaluation experiments across multiple different datasets. The last one is "Combating Label Noise in Deep Learning Using Abstention" (Sunil Thulasidasan, 2019). The author introduces a novel method to combat label noise when training deep

neural networks for classification. Their method gives us the clue to combating the different type of label noises into the training dataset.

4 Problem Statement

After reviewing all related works above and summarizing the key points, we decided to conduct a survey that examine the effect of label noise on Selective Prediction/Classification. For our experiments, We mainly conduct the 2 tasks, the image classification and the semantic segmentation, in the computer vision domain in terms of different amount of label noise, different scales of label spaces, and across multiple datasets. And we are training models with label noise in the training dataset, not like some of the paper simply use Selective Prediction techniques to do the OOD detection which apply the label noise in the test dataset.

4.1 Selective Prediction & MaxProb

For classificiation and segmentation MaxProb is used in the selective prediction. The softmax creates a probability distribution of each class. Maxprob is defined as the argmax of softmax response. The selection classifier (El-Yaniv and Wiener, 2010b) is two functions (f,g) where f is the classifier and g is the selection function. The highest percentage of the distribution determines the class; however, the percentage alongside the coverage will determine if the models refrain from choosing.

$$(f,g)(x) := \begin{cases} f(x) & g(x) = 1; \\ refrain & g(x) = 0. \end{cases} \quad (1)$$

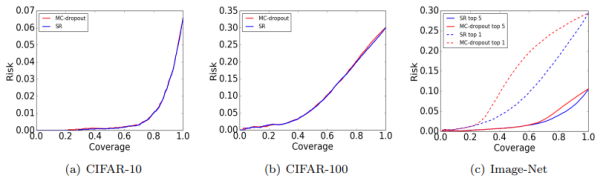
4.2 Risk-Coverage Curve AURC

We propose to use the Area Under Risk-Coverage Curve (AURC) (Yukun Ding, 2020) as the evaluation metric since it is an effective metric to measure performance of selective prediction (Yonatan Geifman, 2019). Risk-Coverage Curve is a curve with coverage as x-axis and risk as y-axis.The coverage denotes the percentage of the input processed by the model without human intervention and the risk denotes the level of risk of these model prediction. Risk is usually defined by a loss function measuring the prediction quality, such as the complement of accuracy. Generally, when models make prediction on more input samples, the coverage becomes larger. At the same time, wrongly predicted samples will increase, and the rate of increase of number of wrongly predicted samples is usually higher than that of truly predicted samples. Hence the risk will also become larger, which will make the Risk-Coverage Curve an increasing function curve. AURC allows us to quantitatively compare the selective prediction results of different trained models. Also, we plot the risk-coverage curves based on softmax response to provide visual aids for analysis. Both ways can help to explain the performance of selective prediction on each model.

5 Experiment Setting

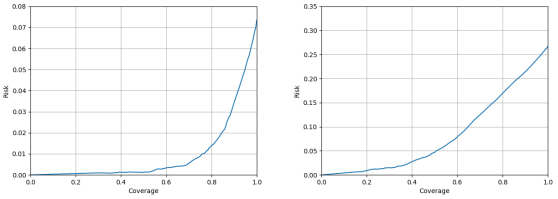
5.1 Baselines and Datasets

For the image classification task, the study of Geifman and El-Yaniv provides the baseline performance of the VGG-16 model on CIFAR-10, CIFAR-100 and ImageNet shown in Figure 1 (Geifman and El-Yaniv, 2017). Due to the limitation

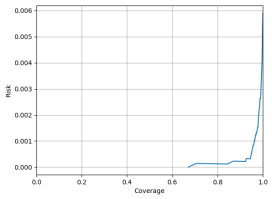


(a) CIFAR-10 (b) CIFAR-100 (c) Image-Net

Figure 1: Risk-coverage curves for (a) cifar-10, (b) cifar-100 and (c) image-net from the study of Geifman and El-Yaniv (Geifman and El-Yaniv, 2017)



(a) CIFAR-10 (b) CIFAR-100



(c) MNIST

Figure 2: Reproduction of risk-coverage curves for VGG-16 on (a) cifar-10, (b) cifar-100 and (c) mnist

of computational resources and time, we reproduce the baseline performance of the VGG-16 model on CIFAR-10 and CIFAR-100. Besides, we create the baseline performance of the same model on

MNIST to diverse the type of experiment datasets. The grayscale MNIST images are converted to RGB format to adapt to the VGG-16 model which takes 3-channel images as input. In terms of experiment settings, model is trained using SGD (Ruder, 2016) with momentum of 0.9, a learning rate of 1.0×10^{-3}, and a weight decay of 5.0×10^{-4}. The loss function is cross entropy loss. With the settings, we manage to reproduce the state-of-the-art performance of VGG-16 on CIFAR-10 and CIFAR-100 with validation accuracies of 92.7% and 73.3% respectively. Also, our trained VGG-16 on MNIST reaches a validation accuracy of 99.41%. The reproduced results shown in Figure 2 are risk-coverage curves and will be used as baselines in our experiments.

For the semantic segmentation task, the FCNs are modified by replacing the VGG-16 backbone model with the pretrained ResNet-18 model (He et al., 2015). To keep the output image size the same as the input size, zero-padding of 100 on each side is applied to the first convolutional layer of the ResNet-18 model. After reconstructing the backbone, we add our own classifier layer including convolutional layers and transpose convolutional layers. Since the Kitti dataset contains 200 images, the dataset is divided into 140 training images, 30 validation images, and 30 test images. Since the majority number of images is of size 375 × 1242 pixel, all images are resized to such image size to enable batch training. The grid search method is used to find the optimal hyperparameters to increase the model performance. Adam optimizer (Kingma and Ba, 2014) and learning rate of 1.0×10^{-4} are used for both models. Learning rate decay of 25% for FCN16 and 10% for FCN8 is applied every 10 epochs. Limited by the GPU memory, batch size of 4 is used for FCN16, and 2 for FCN8. The mean IoU of the trained model is 0.3602 for FCN16, and 0.4176 for FCN8. The result risk-coverage curve is shown in Figure 3. These results will be used as baselines for semantic segmentation task in our experiments.

5.2 Type of noise labels

For the image classification task, VGG-16 models with various label noise are trained and finetuned using the same experiment settings as the baseline model, such as SGD with momentum and weight decay. To generate noisy labeled training data, two types of label noise are used. The type-I label noise

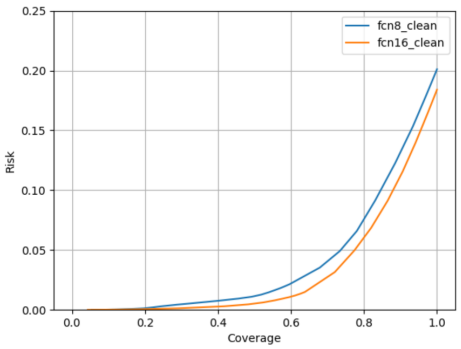


Figure 3: Risk-coverage curves for FCN16 and FCN8 on Kitti dataset

is the uniform noise (Görkem Algan, 2020). For example, given the training images, an noise factor α ratio of the images labeled with i will be changed to uniformly distributed label $j \neq i$. The type-II label noise is the asymmetric noise. For example, an α ratio of the images labeled with i will be changed to $j = i+1$. The performance of selective classification will be evaluated with the two types of label noise with different α. These two methods introduce random noises into the labels, in different ways.

For the semantic segmentation task, we proposed a more realistic and semantic method to introduce noises into the labels. For each pixel in the image, the majority of the labels of the k×k nearest neighbor pixels is used to replace the original label. Take the example in Figure 4, among the nearest 5 by 5 neighbors of the central pixel, there are 12 yellow pixels, 8 green pixels and 5 blue pixels. Therefore after the pollution operation which involves noises into the labels, the label of the central pixel will be yellow, while its original and true label is green. This type of noises is named boundary noise by us, because as shown in Figure 5, most of the noise labels will be located at the boundaries of multiple objects, while the labels inside objects will be kept clean.

6 Results & Discussion

6.1 Image classification

For the image classification task, VGG-16 results are shown in Figure 6, Figure 7, and Figure 8 after applying label noise to each dataset. For MNIST dataset with type-I noise, the risk-coverage curve does not depart much from the baseline until α gets over about 0.7. It indicates that the reasonable amount of type-I label noise may not influence the performance of selective classification on a nearly

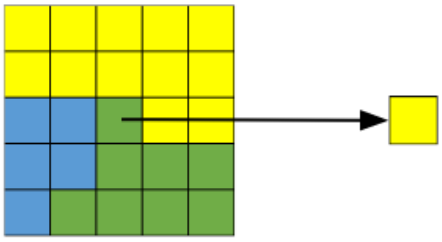
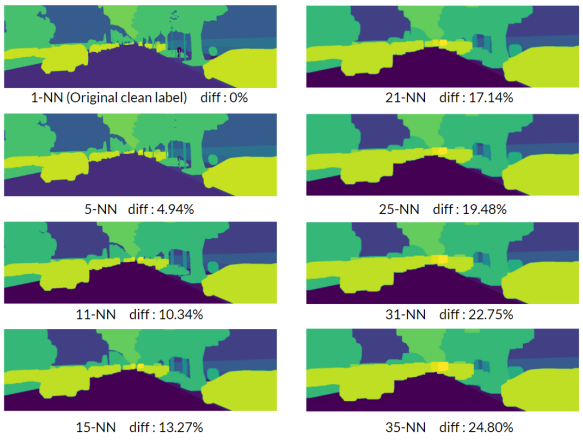


Figure 4: Example of the method introducing boundary noises. Among the nearest 5 by 5 neighbors of the central pixel, there are 12 yellow pixels, 8 green pixels and 5 blue pixels. Therefore after the pollution operation which involves noises into the labels, the label of the central pixel will be yellow, while its original and true label is green.



(a)

(b)

Figure 5: Example of a Kitti image with boundary noises (a) with different neighbor sizes, diff means the portion of pixels changed labels by the method (b) difference between 35-NN image and original clean image, marked in black

solved dataset. Such results can also be verified quantitatively in Figure 6 (c). The AURC curve increases by a small amount when α grows.

However, such results do not apply with different types of the label noise and across datasets. In terms of type of the label noise, applying type-II noise to MNIST dataset has a devastating effect on accuracy and selective prediction when $\alpha \geq 0.5$. The result is reasonable because the dominance ratio of the correct label on each label during the training would decay faster with the increase of α. For instance, training with over 50% amount of type-II noise makes the model tend to classify image to be $i+1$ instead of i. Therefore, type-II label noise generally worsen the performance of the model and selective classification. Regarding the datasets, both types of label noise worsen the performance of selective classification on CIFAR-10 and CIFAR-100 datasets. One main reason of the negative effect can be that the full-coverage risk itself being higher with label noise. Another reason may be that adding more label noise will introduce the model with more degree of uncertainty. Since we are using softmax response as the threshold to generate risk-coverage trade-off, the uncertainty in the model will possibly worsen the performance of selective classification. Furthermore, considering the label space, CIFAR-100 has a larger label space than CIFAR-10. The effect of label noise on selective classification seems to be independent of the label space since the AURC curve both increases by a certain amount with the two CIFAR datasets.

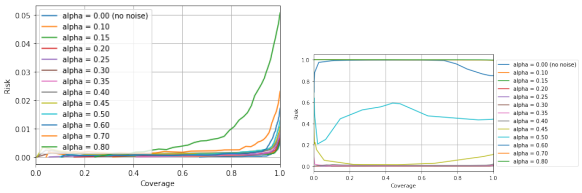
For the ResNet18, results are shown in Figure 9, Figure 10, and Figure 11.

For the ResNet50, results are shown in Figure 12, Figure 13, and Figure 14.

The effect of the type-I and type-II label noise applied on ResNet18 and ResNet50 are very similar as applied on VGG16. However, they still have the different performance.

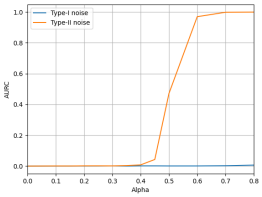
In Figure 15, although ResNet18 does not shows a much difference in SP results comparing to VGG, the overall performance of ResNet18 is still better than VGG.

In Figure 16, ResNet50 shows a much better SP results as we see here the in the bottom 2 plots (c) and (d), the curves are more compact. The reason may be that ResNet50 is a model with more capability, leading to better SP performance.



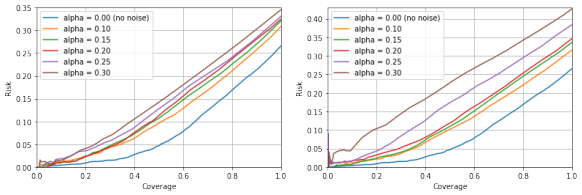
(a) Risk-coverage for Uniform Noise (type-I)

(b) Risk-coverage for Asymmetric Noise (type-II)



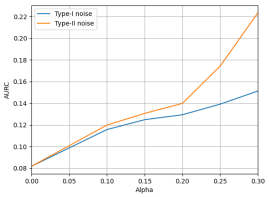
(c) AURC

Figure 6: Experiment Results - VGG16 on MNIST for (a) Risk-coverage for Uniform Noise (type-I), (b) Risk-coverage for Asymmetric Noise (type-II) and (c) AURC



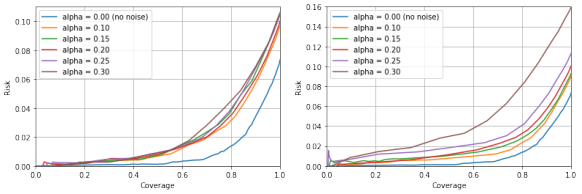
(a) Risk-coverage for Uniform Noise (type-I)

(b) Risk-coverage for Asymmetric Noise (type-II)



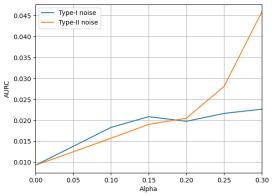
(c) AURC

Figure 8: Experiment Results - VGG16 on CIFAR100 for (a) Risk-coverage for Uniform Noise (type-I), (b) Risk-coverage for Asymmetric Noise (type-II) and (c) AURC



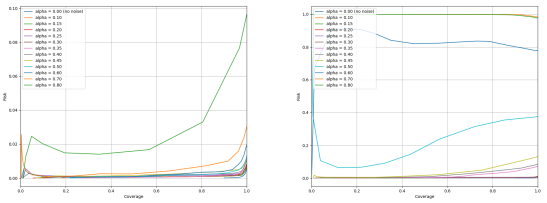
(a) Risk-coverage for Uniform Noise (type-I)

(b) Risk-coverage for Asymmetric Noise (type-II)



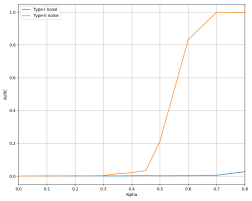
(c) AURC

Figure 7: Experiment Results - VGG16 on CIFAR10 for (a) Risk-coverage for Uniform Noise (type-I), (b) Risk-coverage for Asymmetric Noise (type-II) and (c) AURC



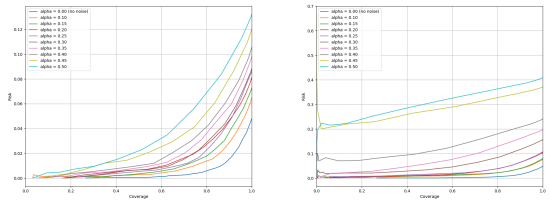
(a) Risk-coverage for Uniform Noise (type-I)

(b) Risk-coverage for Asymmetric Noise (type-II)



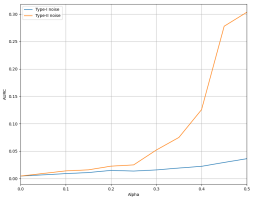
(c) AURC

Figure 9: Experiment Results - Resnet18 on MNIST for (a) Risk-coverage for Uniform Noise (type-I), (b) Risk-coverage for Asymmetric Noise (type-II) and (c) AURC



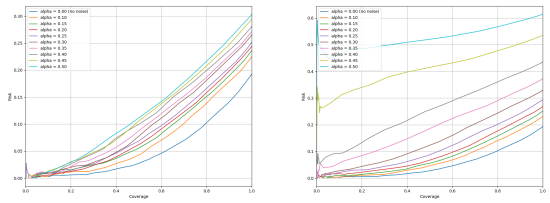
(a) Risk-coverage for Uniform Noise (type-I)

(b) Risk-coverage for Asymmetric Noise (type-II)



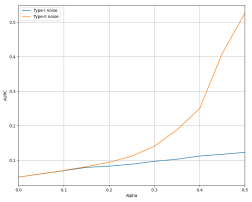
(c) AURC

Figure 10: Experiment Results - Resnet18 on CIFAR10 for (a) Risk-coverage for Uniform Noise (type-I), (b) Risk-coverage for Asymmetric Noise (type-II) and (c) AURC



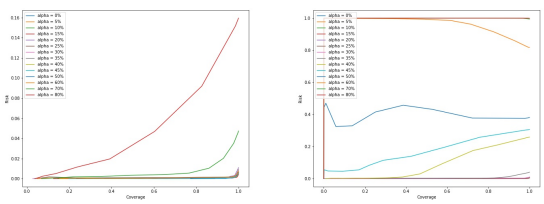
(a) Risk-coverage for Uniform Noise (type-I)

(b) Risk-coverage for Asymmetric Noise (type-II)



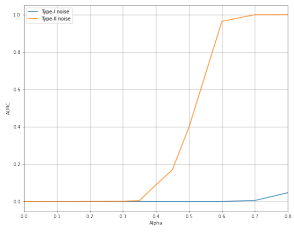
(c) AURC

Figure 11: Experiment Results - Resnet18 on CIFAR10 for (a) Risk-coverage for Uniform Noise (type-I), (b) Risk-coverage for Asymmetric Noise (type-II) and (c) AURC



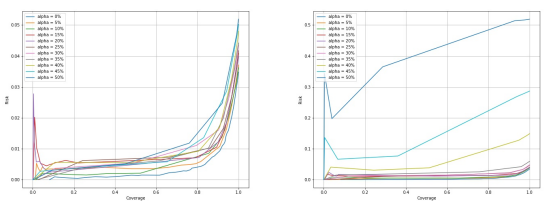
(a) Risk-coverage for Uniform Noise (type-I)

(b) Risk-coverage for Asymmetric Noise (type-II)



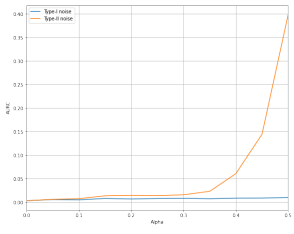
(c) AURC

Figure 12: Experiment Results - Resnet18 on MNIST for (a) Risk-coverage for Uniform Noise (type-I), (b) Risk-coverage for Asymmetric Noise (type-II) and (c) AURC



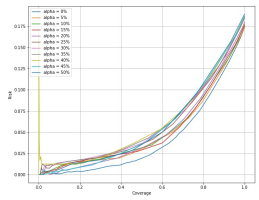
(a) Risk-coverage for Uniform Noise (type-I)

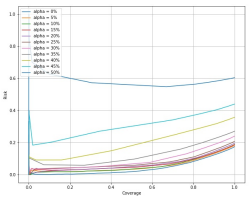
(b) Risk-coverage for Asymmetric Noise (type-II)



(c) AURC

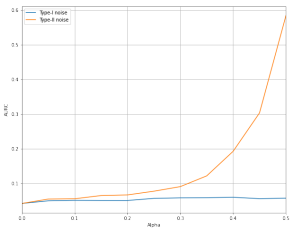
Figure 13: Experiment Results - Resnet18 on CIFAR10 for (a) Risk-coverage for Uniform Noise (type-I), (b) Risk-coverage for Asymmetric Noise (type-II) and (c) AURC





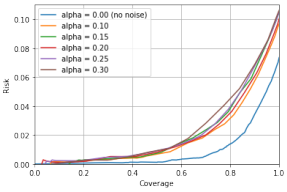
(a) Risk-coverage for Uniform Noise (type-I)

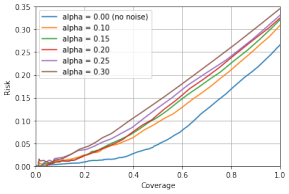
(b) Risk-coverage for Asymmetric Noise (type-II)



(c) AURC

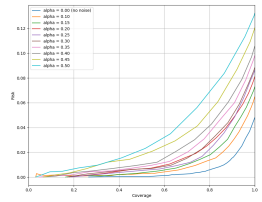
Figure 14: Experiment Results - Resnet18 on CIFAR10 for (a) Risk-coverage for Uniform Noise (type-I), (b) Risk-coverage for Asymmetric Noise (type-II) and (c) AURC

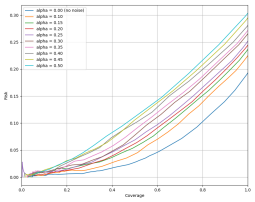




(a) VGG16 on CIFAR10

(b) Risk-coverage for Asymmetric Noise (type-II)

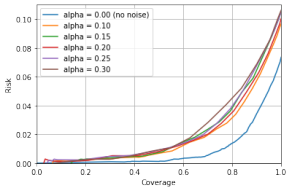


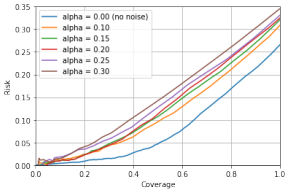


(c) ResNet18 on CIFAR10

(d) ResNet18 on CIFAR100

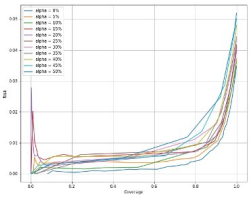
Figure 15: Experiment Results - VGG16 vs. ResNet50 for (a) VGG16 on CIFAR10, (b) VGG16 on CIFAR100, (c) ResNet50 on CIFAR10 and (d) ResNet50 on CIFAR100

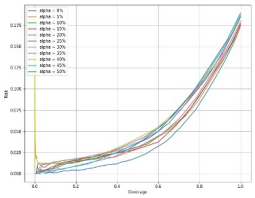




(a) VGG16 on CIFAR10

(b) Risk-coverage for Asymmetric Noise (type-II)





(c) ResNet50 on CIFAR10

(d) ResNet50 on CIFAR100

Figure 16: Experiment Results - VGG16 vs. ResNet50 for (a) VGG16 on CIFAR10, (b) VGG16 on CIFAR100, (c) ResNet50 on CIFAR10 and (d) ResNet50 on CIFAR100

6.2 Semantic segmentation

We tested the Softmax Response method with two fully convolution network backbones with different capabilities, FCN8 and FCN16. We introduced boundary noises at different scales, from 5-NN to 35-NN. The reference noise rate is from 4.94% to 24.80%, note that the noise rates are for reference only, because the distribution of noises is not random.

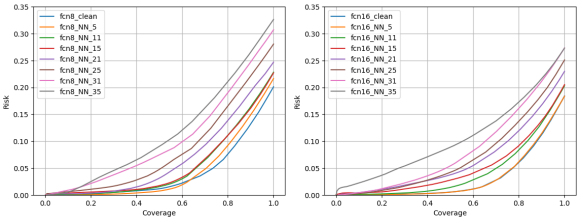
The results is shown in Table 1 and Figure 17. We recorded and calculated the increments of AURC (for FCNs with selective prediction) and risk (for FCNs without selective prediction). As more noises are introduced in the labels, both AURC and risk increased in two models, which is intuitive. And we can compare the increment and get some interesting results. Comparing models with selective prediction (FCN8-SP and FCN16-SP), we can see that the absolute increment is very close, but FCN16-SP has a higher relative increment. On the contrary, comparing models without selective prediction (FCN8 and FCN16), although the absolute increment is still close, but we can see that FCN8 has a higher relative increment. This may indicate that models with and without selective prediction may be affected differently by noise labels, although using the same backbone model architecture.

Another observation is that, compared to models

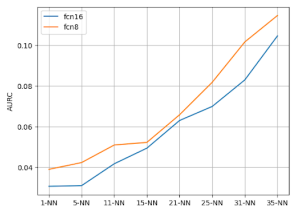
	FCN8-SP	FCN16-SP	FCN8	FCN16
	AURC		risk	
1-NN	0.0390 (100.0%)	0.0306 (100.0%)	0.2011 (100.0%)	0.1840 (100.0%)
5-NN	0.0423 (108.6%)	0.0309 (101.1%)	0.2164 (107.6%)	0.1835 (99.8%)
11-NN	0.0509 (130.7%)	0.0417 (136.2%)	0.2282 (113.5%)	0.2029 (110.3%)
15-NN	0.0522 (134.0%)	0.0494 (161.3%)	0.2269 (112.9%)	0.2051 (111.4%)
21-NN	0.0658 (168.9%)	0.0630 (205.7%)	0.2469 (122.8%)	0.2298 (124.9%)
25-NN	0.0817 (209.9%)	0.0699 (228.2%)	0.2805 (139.5%)	0.2510 (136.4%)
31-NN	0.1016 (260.9%)	0.0829 (270.9%)	0.3071 (152.7%)	0.2733 (148.6%)
35-NN	0.1146 (294.2%)	0.1046 (341.6%)	0.3262 (162.2%)	0.2731 (148.4%)

Table 1: AURC for semantic segmentation experiment

without selective prediction, models with selective prediction is more susceptible to noisy labels. We can see that in the 35-NN case, FCN-SPs take about 300% risk, while FCNs only take about 150% risk. This may guide a direction for future work. The research on selective prediction should pay more attention to the influence of noises in the datasets. At the same time, it is hoped that more people can focus on introducing a module to overcome the affect of noise labels in the selective prediction model. Another little discovery is that, in the 5-NN case, both FCN8-SP and FCN16-SP have little increment on AURC. This may indicate that selective prediction is resistant to a small amount of noises, therefore in the future work, only experiments that introduce more noises are valuable for research.



(a) Risk-coverage for FCN8 (b) Risk-coverage for FCN16



(c) AURC

Figure 17: Experiment Results - Semantic segmentation with boundary noise for (a) Risk-coverage for FCN8, (b) Risk-coverage for FCN16 and (c) AURC

6.3 Limitation

As the focus of the project is to see the effects of label noise on selective prediction. The models were not fully optimized as a generally well-trained model would be sufficient to display the effects. The models were limited to VGG and ResNet. And due to the constraints on time and computing power, the experiments are only done with one selective prediction method, softmax response. Although it is a top performer method, research comparing different selective prediction methods is still meaningful and expected. And for semantic segmentation, due to the same reason, experiments on different datasets and types of noise labels are reserved for future work. Label noise could affect selective prediction differently on other models.

7 Conclusion

In conclusion, for image classification task, the label noise may negatively impact the performance of selective prediction. A harder dataset, a larger label space, and a less capable DNN model will likely result in worse selective prediction performance with label noise. For the semantic segmentation task, noise labels will also cause negative effects. And we can conclude that models with selective prediction is more sensitive to label noises than models without selective prediction. And we see that models with different capabilities will be affected in different manner. This may inspire future works that, the research on selective prediction with noise labels should be performed under as many different conditions as possible. The conditions include models with different capabilities, different datasets, different selective prediction methods, and different types of noises.

8 Teamwork Partition

Jun Chen works on the experiments of image classification with ResNet50 backbone. Yitong Chen works on the experiments of image classification with VGG16 and ResNet18 backbones. Yifan Zheng works on the experiments of semantic segmentation. Yuhan Zhao works on the report drafts. Overall, we distribute the work evenly.

References

Hassan Alhaija, Siva Mustikovela, Lars Mescheder, Andreas Geiger, and Carsten Rother. 2018. Augmented reality meets computer vision: Efficient data generation for urban driving scenes. *International Journal of Computer Vision (IJCV)*.

Ran El-Yaniv and Yair Wiener. 2010a. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641.

Ran El-Yaniv and Yair Wiener. 2010b. On the foundations of noise-free selective classification. volume Journal of Machine Learning Research, 11, Article 1605-1641.

Sang Michael Xie Percy Liang Fahim Tajwar, Ananya Kumar. 2021. No true state-of-the-art? ood detection methods are inconsistent across datasets. volume arXiv:2109.05554.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. volume arXiv:1705.08500.

İlkay Ulusoy Görkem Algan. 2020. Label noise types and their effects on deep learning. volume arXiv:2003.10471.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Andrew Zisserman Karen Simonyan. 2015. Very deep convolutional networks for large-scale image recognition. volume arXiv:1409.1556.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. Cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images.

Yann LeCun and Corinna Cortes. 2010. MNIST handwritten digit database.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2014. Fully convolutional networks for semantic segmentation. *CoRR*, abs/1411.4038.

Sebastian Ruder. 2016. An overview of gradient descent optimization algorithms. *arXiv preprint*, arXiv:1609.04747.

Jeff Bilmes Gopinath Chennupati Jamal Mohd-Yusof Sunil Thulasidasan, Tanmoy Bhattacharya. 2019. Combating label noise in deep learning using abstention. volume arXiv:1905.10964.

Neeraj Varshney, Swaroop Mishra, and Chitta Baral. 2022. Investigating selective prediction approaches across several tasks in iid, ood, and adversarial settings. volume arXiv:2203.00211.

Ran El-Yaniv Yonatan Geifman, Guy Uziel. 2019. Bias-reduced uncertainty estimation for deep neural classifiers. volume arXiv:1805.08206.

Jinjun Xiong-Yiyu Shi Yukun Ding, Jinglan Liu. 2020. Revisiting the evaluation of uncertainty estimation and its application to explore model complexity-uncertainty trade-off. volume arXiv:1903.02050.