Effect of Label Noise on Robustness of Deep Neural Network Object Detectors

Bishwo Adhikari$^{1[0000-0003-0037-8313]}$,
Jukka Peltomäki$^{1[0000-0002-9779-6804]}$,
Saeed Bakhshi Germi$^{1[0000-0003-3048-220X]}$,
Esa Rahtu$^{1[0000-0001-8767-0864]}$, and
Heikki Huttunen$^{2[0000-0002-6571-0797]}$

[1] Tampere University, Tampere, Finland
[2] Visy Oy, Tampere, Finland
{bishwo.adhikari, jukka.peltomaki, saeed.bakhshigermi, esa.rahtu}@tuni.fi
heikki.huttunen@visy.fi

Abstract. Label noise is a primary point of interest for safety concerns in previous works as it affects the robustness of a machine learning system by a considerable amount. This paper studies the sensitivity of object detection loss functions to label noise in bounding box detection tasks. Although label noise has been widely studied in the classification context, less attention is paid to its effect on object detection. We characterize different types of label noise and concentrate on the most common type of annotation error, which is missing labels. We simulate missing labels by deliberately removing bounding boxes at training time and study its effect on different deep learning object detection architectures and their loss functions. Our primary focus is on comparing two particular loss functions: cross-entropy loss and focal loss. We also experiment on the effect of different focal loss hyperparameter values with varying amounts of noise in the datasets and discover that even up to 50% missing labels can be tolerated with an appropriate selection of hyperparameters. The results suggest that focal loss is more sensitive to label noise, but increasing the gamma value can boost its robustness.

Keywords: Safe AI · Deep Neural Networks · Label Noise · Image Labeling

1 Introduction

The growing success of deep neural network algorithms in solving challenging tasks resulted in a surge of interest from the safety-critical applications domain. As stated by recent works, one of the major issues of using such an algorithm in line with safety standards is the effects of label noise on the output [1–3].

Earlier object detection pipelines consisted of manually engineered feature extraction together with relatively simple classifiers [4,5]. These systems required a human to label the different objects for training, and the labeling was done on the crop level. Although this approach had its challenges, such as mining negative examples, its behavior is still reasonably well understood due to relying on a straightforward method.

More recently, the success of convolutional neural networks (CNN) and deep learning [6] has transformed the domain of object detection. These approaches outperform traditional techniques by a large margin but are also more data-hungry at the same time [7–9]. The tedious task of manual labeling of enormous datasets means there will be faults in the process inevitably.

Popular large object detection datasets include MS COCO [10], PASCAL VOC [11], and OpenImages [12], containing millions of examples with quality annotations. The ground truth human annotations are gathered by crowdsourcing, and elaborate reward and evaluation schemes guarantee high quality for the annotations. However, apart from these large annotation campaigns, many players, companies, and research groups routinely collect smaller datasets within their application domains. In such cases, the quality of annotations is often compromised due to limited resources. Moreover, even standard benchmark sets are not error-free, and the influence of erroneous annotations on the system's safety requires further study.

The presence of noise in the training dataset can have a severe impact on the system's performance. For example, in the video surveillance system, a good detector would retain the same confidence, box coordinates, and class label over time. On the other hand, a bad one will be flickering, where the confidence fluctuates, the coordinates change, and even the class is mislabeled from frame to frame.

Figure 1 shows the four most common annotation error categories found in object detection datasets. These categories are (a) missing annotations (false negatives), (b) extra annotations (false positives), (c) inaccurate bounding boxes (which would result in low intersection over union (IoU)), and (d) incorrect class labels. In our experience, the most common error type is the first one, where the human annotator misses some target objects due to occlusions, small size, a large number of objects, or simply unclear annotation instructions. The second most frequent annotation error type is inaccurate bounding boxes, a very natural error for a human, as it takes more time and effort to pay attention to detail in every case. The two other types in Figure 1, completely incorrect annotations and wrong labels, are probably easier for humans to avoid.

The loss functions being a significant differentiator in modern single-stage detection pipelines and current challenges for annotation quality, inspired us to study the effect of label noise in object detection with two popular loss functions. Notably, in this paper, we focus on examining how *cross-entropy loss* (CE) and *focal loss functions* (FL) handle noise in the form of missing labels. We focus on these losses since the focal loss is commonly used but may suffer from missing annotations because it puts higher weight on complex samples (hard negatives and hard positives). Missing bounding boxes in the annotation appear as hard negatives from the training point of view, and we wish to study their influence on the resulting accuracy. The main contributions of this paper are:

- We characterize different types of noise present in object detection datasets.
- We provide empirical observations on training single-stage object detectors with different loss functions and different hyperparameter settings.
- We suggest possible measures to boost the robustness of the object detector with minimal changes in the network.

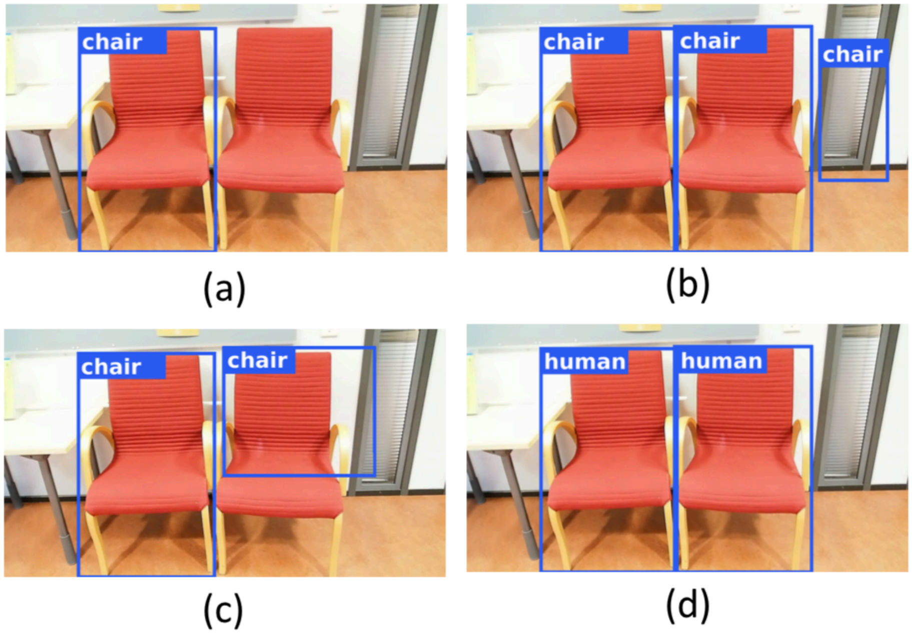

Fig. 1. Common types of label noise in object detection. (a) Missing label, the other chair is not labeled. (b) Incorrect annotation. (c) Inaccurately drawn box, resulting in low IoU. (d) Wrong classification label, humans instead of chairs. Image from the Indoor dataset [13].

The remainder of this paper is structured as follows. Section 2 briefly summarizes the related works followed by the review of object detection loss functions in Section 3. In Section 4, we experimented with multiple scenarios on our hypothesis and analyzed the obtained results. Finally, we conclude this paper with our findings and future direction in Section 5.

2 Related Works

Willers [1] and Wozniak [2] both provide a list of safety concerns or goals related to deep learning algorithms. In their works, label noise is mentioned as one of the primary faults that can affect safety. It is suggested to have a labeling guideline to mitigate the effect of this fault. However, even with a guideline, manually labeling a large dataset is prone to noise, as discussed before. Thus, a proper approach is required to deal with noisy datasets in deep learning systems. Zhang [14] reviews problems related to the dataset, such as label noise, by surveying over recent works. According to his work, using a robust loss function and reweighting samples can help mitigate this issue.

Our topic of label noise in object detection is closely related to the topic of label noise in image classification, which has been studied more: For image classification, Frenay and Verleysen [15] have proposed a taxonomy of different types of noise, studied their consequences, and reviewed multiple techniques to clean noise and have the algorithms be more noise-tolerant. Li *et al.* have proposed BundleNet exploiting sample correlations by creating bundles of samples class-by-class and treating them as independent inputs, which acts as a noise-robust regularization mechanism [16]. Lee *et al.* have proposed CleanNet to detect noise in the dataset and be used in tandem with a classifier network for better noise tolerance [17].

Noise in object detection is different from classification because an image can have any natural number of objects present, anywhere in the image. A label in object detection is a box with a position, a size, and a class, which adds more possibilities for noise. It is easier for a human annotator to identify that an object in a picture is indeed a banana than correctly labeling dozens of bananas in one image of a cafeteria. The tedious task of doing so might result in the human annotator skipping some labels. Skipping a label causes label noise in the form of a missing label. Moreover, the task is often ambiguous when dealing with objects in a real-world image. Partially occluded objects, reflective surfaces, distance to the camera, and overcrowded images become relevant consideration points when labeling for object detection. These problems make the human annotator's role more prominent because more mental decisions are required. It also means that there will be more variation in the annotations, as different humans make different decisions.

Su *et al.* [18] have studied the overall process of annotation for object detection in a crowd-sourced manner. They first divided the task into three different sub-tasks: (1.) draw a box, (2.) verify the quality of a drawn box, and (3.) verify a box coverage on a single image. Different people do all these sub-tasks via Amazon Mechanical Turk (AMT). They concluded that this method produces good quality annotations.

Russakovsky *et al.* [19] have studied the human-in-the-loop annotation process, where state-of-the-art object detection models are used to detect many of the objects in the image. Then humans are used for detecting all the objects that the models are unable to detect. This method is needed as no current object detection system is perfect, yet, and their goal is to have every object in the image annotated adequately. A properly annotated object should have a tightly fitted box and not an arbitrary margin of non-object space in the annotation. They conclude that their method of using humans and computer vision together was better than using either alone.

3 Object Detection Loss Functions

Single-shot detection (SSD) [8] uses both regression loss for bounding box regression and cross-entropy loss for classification. The cross-entropy loss for a sample with ground truth one-hot-encoded labels $\mathbf{y} = (y_1, y_2, \ldots, y_C)$ and predicted class confidences $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_C)$ in a C-class classification problem is defined as

$$\mathrm{CE}(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{c=1}^{C} y_c \log(\hat{y}_c). \tag{1}$$

The focal loss was extended by Lin *et al.* [20] to handle difficult samples better. They show that this improvement can result in better accuracy compared to the cross-entropy loss. The focal loss was designed to emphasize hard positives. It is similar to cross-entropy loss but has a parameterized penalty factor $\gamma > 0$ weighing the influence of each sample based on its detection score. More specifically, the focal loss for the C-class classification with ground truth $\mathbf{y} = (y_1, y_2, \ldots, y_C)$ and predictions $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_C)$ is defined as

$$\mathrm{FL}(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{c=1}^{C} \alpha_c (1 - \hat{y}_c)^{\gamma} y_c \log(\hat{y}_c), \tag{2}$$

with the balancing factor α_c [20], which is equal to 0.75 for all $c \in \{1, \ldots, C\}$ in all our experiments.

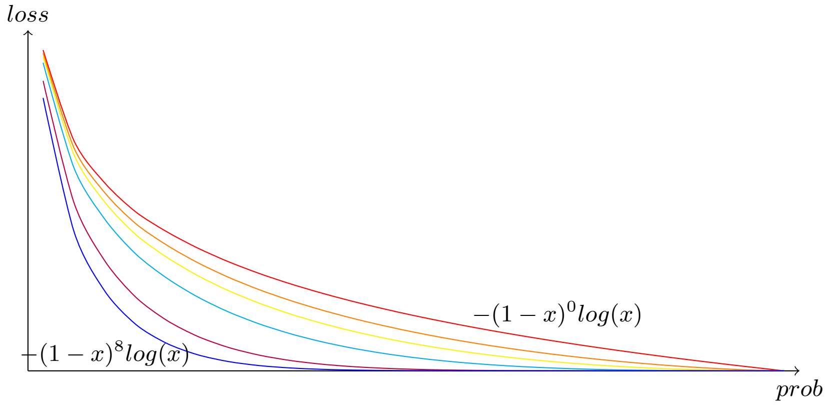

Fig. 2. Visualization of the focal loss function with different values for parameter $\gamma = 0, 1, 2, \ldots, 8$. The probability of being the ground truth is on the horizontal axis, and the loss is on the vertical axis. The higher the gamma value, the sharper the focus on harder cases. With gamma equaling zero, the focal loss is the same as cross-entropy loss.

In other words, the FL loss differs from the CE loss by the weight term $(1-\hat{y}_c)$, whose effect is to assign a higher weight for samples with low confidence (small

$\hat{y}_c$). γ affects the overall loss by lowering it; primarily well-classified samples with high confidence y_c for the most likely class c will have a negligible loss. At the same time, more attention is paid to learning the more complicated cases. Figure 2 demonstrates this scaling and aptly visualizes how the different γ parameters change the ferocity of the focus on more complicated cases. However, this loss weighting may have an adverse effect in the presence of label noise. The missing annotations are viewed as hard positives (non-annotated targets found by the model with a nonzero likelihood).

4 Experiments and Results

It was observed that sometimes in custom datasets, the focal loss seemed to produce results that were not as good as the research suggested. The intuition was formed that the weighting of complex cases, as performed by the focal loss function, would be more sensitive to label noise. The reason is that if a label is erroneous, to begin with, it is impossible to get right, so focusing on such a label leads the model astray and misuses the model capacity.

The experiments consider two questions: (1) how does label noise affect the two losses, and (2) how do models trained with different γ values tolerate label noise. For both experiments, we study the performance with three datasets: first with a small high-quality Indoor dataset, the second uses a large classical PASCAL VOC dataset, which does contain some annotation errors natively, and finally, with a single class FDDB dataset. Table 1 contains the characteristics of these datasets.

In all our experiments, the single-stage object detection (SSD) with MobileNet v1 [21] backbone network is fine-tuned from MS COCO pre-trained model for 100K training steps. We experimented only with the missing labels category. So, the training dataset has a percentage of randomly missing annotation boxes.

Table 1. Comparison of Indoor [13], PASCAL VOC 2012 [11] and FDDB [22] datasets based on source, size, quality of annotation, and usages.

	Indoor	**PASCAL VOC**	**FDDB**
Sample Source	Indoor scenes	Collected online	Faces in the Wild
Image Count	2213	17125	2845
Amount of Instances	4500	40000	5171
Number of Classes	7	20	1
Usage	Object detection	Multi-purpose	Face detection

4.1 Noise robustness of the two losses: CE vs. FL

In this experiment, we use six different noise levels: 0%, 10%, 20%, 30%, 40% or 50%, of missing labels. The dropping of the labels was done randomly, but both networks were using the same training datasets. Also, the noisy datasets are constructed incrementally, *i.e.,*, the 20% noise had all the labels of the 10% dataset dropped (+10% more), and so forth. The model with both the CE loss and the FL loss with hyperparameter $\gamma = 2$ (as proposed in the original paper [20]) is fine-tuned for 100K steps, and *mAP@.50IoU* (mean average precision with 0.5 IoU threshold) is used as a performance evaluation metric.

Indoor dataset— In the first set of experiments, we start training SSD using pre-trained weight from the MS COCO dataset, where some classes overlap between the datasets (chair, TV set, . . .), while others do not (fire extinguisher). The resulting accuracies are presented in Figure 3a; mAP@0.50 with the CE loss and the FL loss. Moreover, we show the *relative drop* in mAP with respect to noiseless labels in Figure 4. It seems that the accuracy resulting from the FL loss objective function outperforms the CE loss for 10% – 20% noise levels. The FL loss is more robust till the 30% noise level and maintains a higher mAP than the CE loss. However, with the higher amount of label noise (> 30%), FL loss accuracy plunged rapidly, falling behind the CE loss. For the extremely noisy (i.e., 50%) training dataset, accuracy from FL loss is 2% lower than that of CE loss.

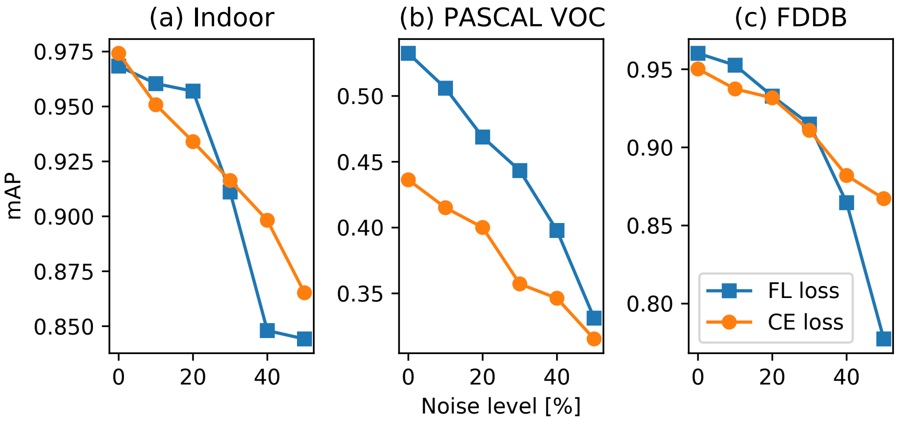

Fig. 3. Relationship between mAP (%) and different amount of noise levels on PASCAL VOC, and FDBB datasets.

PASCAL VOC dataset— Next, we studied the noise sensitivity on the PASCAL VOC [11] dataset. The network using the FL loss function performs better than the alternative, but the accuracy with FL loss decreases more when

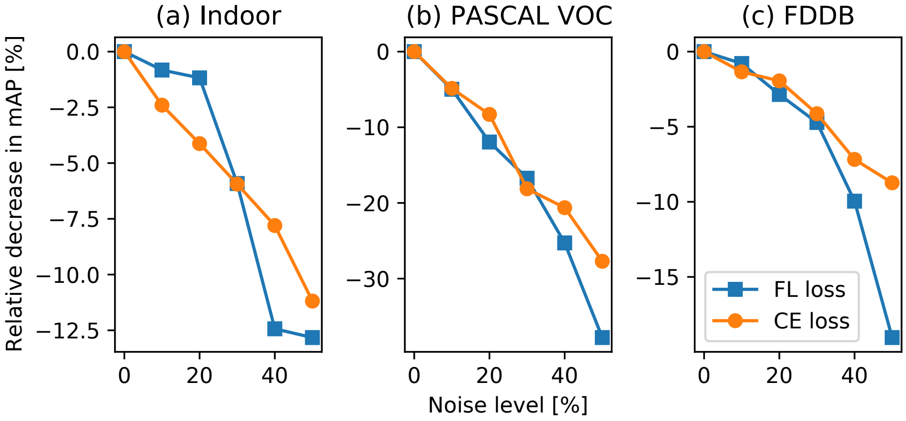

Fig. 4. Relative decrease in mAP (%) with respect to the noise levels in Indoor, PASCAL VOC, and FDDB datasets.

the noise level increases compared to CE loss. Without added noise, FL loss gives 10% higher mAP than CE loss. While the FL loss outperforms the CE loss in detection performance, it has a higher rate of mAP decrease than the CE loss. The difference in detection performance gets smaller by increasing label noise. The drop in mAP from no added noise to 50% label noise is 20.12% in with FL loss and 12.10% with CE loss.

FDDB dataset— Next, we studied the noise sensitivity on the high-quality moderate-sized single class dataset, Face Detection Data Set and Benchmark (FDDB)[22]. As shown in Figure 2c, the network using the FL loss function performs better till the 30% noise label. Adding more noise to the training dataset causes the accuracy to drop. The performance difference is smaller for lower noise levels and gets more significant for the noisy cases. The drop in AP from no added noise to 50% label noise is 18.28% in the FL loss case and 8.03% CE loss case.

Overall, the two losses seem to have similar behavior with these datasets. Compared to the FL loss, the CE loss is *more* robust to increased noise levels. However, with the VOC dataset, even though the FL loss suffers more for extreme cases, the overall performance remains higher than the CE loss at all points shown in Figure 2b.

We speculate that the Indoor and FDDB are relatively easy compared to VOC, containing fewer small (difficult) bounding boxes. Thus, as long as most bounding boxes are in place, the FL loss equally weights the true targets and the hard negatives produced by the missing labels. The more varied and challenging nature of the PASCAL VOC dataset causes different noise tolerance behaviors than the smaller datasets.

4.2 Effect of the gamma parameter (γ)

In our second set of experiments, we compare the robustness of the FL loss for different values of the γ parameter. This time we only ran for three noise levels: 0%, 10%, and 50%. The gamma values tested were $\gamma = 1, 2, \ldots, 8$. All the other settings were kept the same as in the previous experiment.

Indoor dataset— The first experiment in this set uses the Indoor dataset; results on this dataset are presented in Figure 5a. In this dataset, the 10% noise detection performance is very close to the 0% noise. More interestingly, with extremely high label noise (50%), the gamma value has a significant impact. With $\gamma = 0$, the accuracy on the clean dataset (0% missing labels) is 18.52% more than the extremely noisy dataset (50% missing labels). With $\gamma = 8$, the clean dataset mAP is only 5.2% higher than the noisy dataset. The mAP curve indicates that a higher γ value does not affect the clean dataset while it boosts the performance in the presence of label noise.

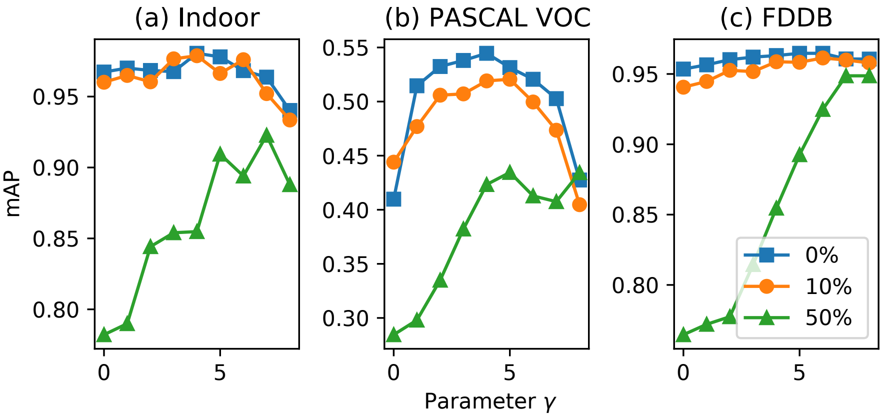

Fig. 5. Results on Indoor, PASCAL VOC and FDDB datasets with different gamma values on 0%, 10% and 50% noise levels.

PASCAL VOC dataset— Next, we experiment on PASCAL VOC with different γ values. The higher values of gamma can be used to offset the effect of missing labels partially. Like the previous experiment, γ values in the range 4 – 6 have better performance.

FDDB dataset— Experiments result on FDDB with different γ values is shown in Figure 5c. Results coincide with our previous experiments. With $\gamma = 0$, the difference in performance between clean and extremely noisy datasets is 18.90%. However, this difference gets smaller by increasing the γ value. With $\gamma = 8$, a clean dataset is only 2% more accurate than a heavily noised dataset.

Generally, with an extreme amount of label noise, increasing the γ value improves the detection results. Still, the exact γ value and the detection performance are dependent on the dataset. This could indicate that maybe the sharp concentration introduced by the higher γ values can offset the missing labels in relatively easy datasets. Experiments on these datasets suggest that the robustness to label noise increases for larger γ values. In these cases, the model essentially learns from the complex samples only (annotated targets detected with low confidence and non-annotated targets detected with high confidence).

This is illustrated in Figure 6, which shows the FL loss curves for both negative and positive examples. Due to the large γ value, the intermediate values ($\hat{y} \in [0.3, 0.7]$) behave as a *don't care* region, and the model does not learn from samples falling into this zone. Since all learning is based on complex samples (similarly to the support vector machine), it will be enough to push all objects with annotations to the "don't care" region. On the other hand, all negative samples (including missing annotations) can safely reside in this zone, and the model essentially learns to ignore those.

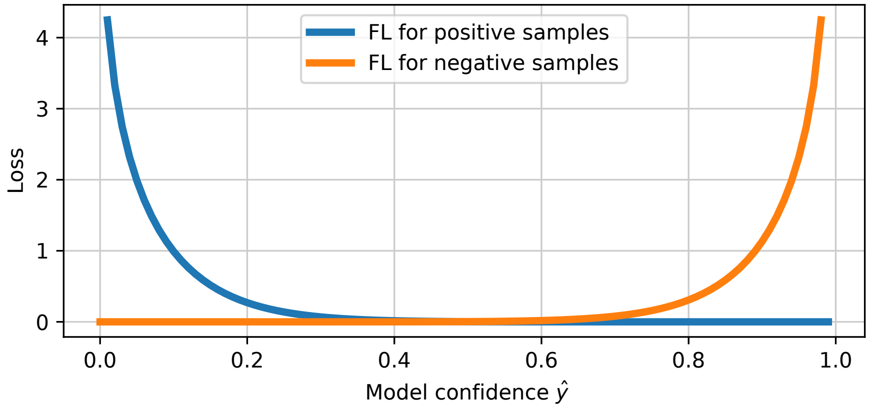

Fig. 6. Focal loss with $\gamma = 8$ for negative and positive samples with respect to model prediction confidence.

5 Conclusion

In this work, we characterized different types of label noise present in object detection datasets and explored the sensitivity of loss functions to them. With label noise being a crucial factor in ensuring the safety of the machine learning algorithm, we made sure to include experiments with large-scale real-world datasets. More specifically, we experimented on three datasets with varying amounts of label noise with cross-entropy and focal loss. Experiments suggest that focal loss suffers more with high amounts of noise, falling behind the cross-entropy loss. The second aspect studied is the effect of the hyperparameter γ on the sensitivity to label noise. It was discovered that larger values of γ improve the robustness to label noise such that extreme gamma values make the model indifferent to the noise level.

For future work, it would be beneficial to run more varied experiments to see how the label noise tolerance differs when training the network from scratch and its effect on system safety. Another point to consider would be running experiments with improved loss functions that are better suited for noisy datasets. It is also possible to quantify the risk associated with mislabeling by taking a statistical approach.

All relevant information, data, and codes are published open-access at `https://github.com/adhikaribishwo/label_noise_on_object_detection`.

Acknowledgment

This work was financially supported by Business Finland project *408/31/2018 MIDAS*.

References

1. Willers, O., Sudholt, S., Raafatnia, S., Abrecht, S.: Safety Concerns and Mitigation Approaches Regarding the Use of Deep Learning in Safety-Critical Perception Tasks. Computer Safety, Reliability, and Security. SAFECOMP Workshops. 12235, 336-350 (2020).
2. Wozniak, E., Cârlan, C., Acar-Celik, E., Putzer, H.: A Safety Case Pattern for Systems with Machine Learning Components. Computer Safety, Reliability, and Security. SAFECOMP Workshops. 12235, 370-382 (2020).
3. Schwalbe, G., Knie, B., Sämann, T., Dobberphul, T., Gauerhof, L., Raafatnia, S., Rocco, V.: Structuring the Safety Argumentation for Deep Neural Network Based Perception in Automotive Applications. Computer Safety, Reliability, and Security. SAFECOMP Workshops. 12235, 383-394 (2020).
4. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). 1, 886-893, (2005).
5. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning. 20, 273-297 (1995).
6. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature. 521, 436-444 (2015).

7. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence. 39, 1137-1149 (2017).
8. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C., Berg, A.: SSD: Single Shot MultiBox Detector. European Conference on Computer Vision (ECCV). 9905, 21-37 (2016).
9. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You Only Look Once: Unified, Real-Time Object Detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 779-788 (2016).
10. Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.: Microsoft COCO: Common Objects in Context. European Conference on Computer Vision (ECCV). 8693, 740-755 (2014).
11. Everingham, M., Eslami, S., Van Gool, L., Williams, C., Winn, J., Zisserman, A.: The Pascal Visual Object Classes Challenge: A Retrospective. International Journal of Computer Vision. 111, 98-136 (2014).
12. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., Ferrari, V.: The Open Images Dataset V4. International Journal of Computer Vision. 128, 1956-1981 (2020).
13. Adhikari, B., Peltomaki, J., Puura, J., Huttunen, H.: Faster Bounding Box Annotation for Object Detection in Indoor Scenes. 7th European Workshop on Visual Information Processing (EUVIP). 1-6 (2018).
14. Zhang, X., Liu, C., Suen, C.: Towards Robust Pattern Recognition: A Review. Proceedings of the IEEE. 108, 894-922 (2020).
15. Frenay, B., Verleysen, M.: Classification in the Presence of Label Noise: A Survey. IEEE Transactions on Neural Networks and Learning Systems. 25, 845-869 (2014).
16. Li, C., Zhang, C., Ding, K., Li, G., Cheng, J., Lu, H.: BundleNet: Learning with Noisy Label via Sample Correlations. IEEE Access. 6, 2367-2377 (2018).
17. Lee, K., He, X., Zhang, L., Yang, L.: CleanNet: Transfer Learning for Scalable Image Classifier Training with Label Noise. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 5447-5456 (2018).
18. Su, H., Deng, J., Fei-Fei, L.: Crowdsourcing Annotations for Visual Object Detection. AAAI Human Computation Workshop. 40-46 (2012).
19. Russakovsky, O., Li, L., Fei-Fei, L.: Best of both worlds: Human-machine collaboration for object annotation. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2121-2131 (2015).
20. Lin, T., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal Loss for Dense Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence. 42, 318-327 (2020).
21. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv (2017).
22. Jain, V., Learned-Miller, E.: FDDB: A Benchmark for Face Detection in Unconstrained Settings. Department of Computer Science, University of Massachusetts. UM-CS-2010-009 (2010).