
Estimating label quality and errors in semantic segmentation data via any model

Vedang Lad^{1,2} Jonas Mueller¹

Abstract

The labor-intensive annotation process of semantic segmentation datasets is often prone to errors, since humans struggle to label every pixel correctly. We study algorithms to automatically detect such annotation errors, in particular methods to score *label quality*, such that the images with the lowest scores are least likely to be correctly labeled. This helps prioritize what data to review in order to ensure a high-quality training/evaluation dataset, which is critical in sensitive applications such as medical imaging and autonomous vehicles. Widely applicable, our label quality scores rely on probabilistic predictions from a trained segmentation model – any model architecture and training procedure can be utilized. Here we study 7 different label quality scoring methods used in conjunction with a DeepLabV3+ or a FPN segmentation model to detect annotation errors in a version of the SYNTHIA dataset. Precision-recall evaluations reveal a score – the soft-minimum of the model-estimated likelihoods of each pixel’s annotated class – that is particularly effective to identify images that are mislabeled, across multiple types of annotation error.

1. Introduction

Semantic segmentation, in which a model classifies each pixel within an image, is a cornerstone task of fine-grained image understanding (Kirillov et al., 2023). To train more effective segmentation models, the size of image datasets has dramatically grown in recent years, especially in fields like radiology, pathology, robotics, and autonomous vehicles. The labeling of semantic segmentation data necessitates pixel-wise annotation of images, which is highly label-intensive and error-prone, thus often failing to produce the “ground truth” it is claimed to be. Training models to output incorrect labels is obviously problematic, but even evaluating models with noisy labels is worrisome in high-stakes

medical/robotics applications built around the segmentation task.

Even significantly simpler to annotate data, e.g. for multi-class/multi-label classification, contain many errors (Northcutt et al., 2021a; Thyagarajan et al., 2023). To address this, there is a growing interest in data-centric algorithms, specifically Label Error Detection (LED) methods, which aim to systematically enhance the quality of these datasets (Kuan & Mueller, 2022; Northcutt et al., 2021b; Muller & Markert, 2019). Here we consider LED in semantic segmentation data, where the prevalence of label errors is likely much greater than in classification data, due to the extra annotation complexity. We consider algorithms to assign a *label quality score* to each image¹, such that the images with the lowest scores are most likely to have some sort of annotation error. These are the images that should be prioritized for review or re-labeling (Wang & Mueller, 2022; Kuan & Mueller, 2022; Goh & Mueller, 2023). An effective score will ensure reviewers do not wastefully inspect correctly-labeled images (high precision) while simultaneously overlooking few mislabeled images (high recall).

In this work, we focus on universal LED methods that can be applied to any segmentation dataset for which someone has trained any standard type of segmentation model. The methods considered here only require predictions from the model and the annotated labels for the images in the dataset, we do not require running inference on additional (synthesized) images (Rottmann & Reese, 2022; Nishi et al., 2021), nor training models in a special nonstandard way (Zhang et al., 2020; Sukhbaatar & Fergus, 2014; Jiang et al., 2018; Zhang & Sabuncu, 2018). This ensures our methods are very easy to apply and widely applicable across domains. More importantly, the best segmentation architectures are constantly evolving, and LED methods that require nonstandard training/inference may no longer be compatible with future state-of-the-art architectures. Since the performance of the segmentation model is obviously important for using it to detect mislabeling, LED methods must remain compatible with the best architectures to provide the best insights about a dataset.

Our experiments study three common types of potential an-

¹Cleanlab ²MIT. Correspondence to: Vedang Lad <vedang@mit.edu>, Jonas Mueller <jonas@cleanlab.ai>.

¹Code to run our method:
<https://github.com/cleanlab/cleanlab>

notation error in a segmentation dataset, where for a specific image, annotators may: (1) overlook an object depicted in the image, i.e. its true segmentation mask has been dropped in the given annotation (*Drop*), (2) select the incorrect class label to apply for an annotated segmentation mask, i.e. its true class has been swapped with another class in the given annotation (*Swap*), or (3) miss some pixels which should have been in/out of the mask when annotating it, i.e. the mask’s true proper location has been shifted in the given annotation (*Shift*). While “ground truth” is often used to refer to the given annotation in the segmentation literature, throughout this paper, we sparingly use “truth” to refer to ideal segmentation masks that a hypothetical perfect annotator would produce. The labels we train a model to predict are merely *noisy* reflections of this underlying truth in most real-world datasets (Northcutt et al., 2021a). Here we present a model-agnostic *softmax* label quality score that is able to accurately detect all 3 types of annotation error, in order to help us avoid training models with untrue data.

2. Related Work

A large body of work has studied robust ML with noisy labels (Song et al., 2022; Wei et al., 2022; Chen et al., 2019; Natarajan et al., 2013), but the importance of *label error detection* alone was only more recently recognized (Muller & Markert, 2019; Northcutt et al., 2021a;b; Kuan & Mueller, 2022). With growing value from algorithms to improve data quality, LED methods have been proposed for more complex supervised learning tasks including multi-label classification (Thyagarajan et al., 2023), sequence prediction in NLP (Wang & Mueller, 2022; Klie et al., 2022), and object detection (Chachuła et al., 2022; Murrugarra-Llerena et al., 2022).

Specifically for semantic segmentation, Vădineanu et al. (2022) demonstrate the destructive effects of annotation errors in medical imaging data. For segmentation in medical imaging, Zhang et al. (2020) introduce a supervised segmentation method for jointly estimating the spatial characteristics of label errors from multiple human annotators. However their approach is model-specific requiring two coupled CNNs trained in a special manner unlike the general label quality scores we consider here. Zhang et al. (2020) also introduce the notion of *under-segmentation*, *over-segmentation*, and *class swaps* as common types of errors in segmentation data, which we also study here.

More similar to our work is the general principle of *error analysis*, in which one sorts images by loss (or other evaluation metric) incurred between the model prediction and annotated label. The worst-predicted images often reveal suspicious annotations or other data anomalies. For segmentation one traditionally uses the likelihood loss or IoU evaluation metric. Both are considered as label quality scores in our study, but each has shortcomings. For LED in segmen-

tation data, Rottmann & Reese (2022) found operating on connected components of each image to be more effective than pixel-level analysis, but such approaches have vastly worse computational complexity than the straightforward softmax label quality score we propose here.

3. Methods for Scoring Label Quality

Semantic segmentation datasets typically consist of many images, each containing a fixed grid of pixels. To label the data, annotators assign each pixel to one of K possible classes (i.e. categories in the segmentation task). For a given image x , a standard semantic segmentation model M generates predicted class probabilities $p = M(x)$, where p_{ijk} represents the estimated probability that the i, j -th pixel in image x belongs to class k . For unbiased evaluation of label quality based on model predictions, we must avoid overfitting. Thus we assume throughout this work that these predicted probabilities are computed *out-of-sample*, derived from a copy of the model that did not encounter x during training. Using any type of model, out-of-sample predictions can be obtained for *every* image in a dataset via say 5-fold cross-validation.

Utilizing p , we initially evaluate the individual fine-grained labels for each pixel. In this context, we adopt effective LED methodologies for standard classification scenarios (Northcutt et al., 2021b; Kuan & Mueller, 2022), treating every pixel as a distinct, independent instance (regardless of the image to which it belongs).

For one image with pixel dimensions $h \times w$, we thus have:

- $\mathbf{p} \in \mathbb{R}^{h \times w \times K}$, where p_{ijk} is the model predicted probability that pixel (i, j) belongs to class k .
- $\mathbf{P} = \{P_{ij}\}$ is the predicted segmentation mask, where $P_{ij} \in \{0, \dots, K - 1\}$ is the predicted class label of pixel (i, j) , i.e. the class with the highest probability = $\underset{k}{\operatorname{argmax}} p_{ijk}$.
- $\mathbf{l} = \{l_{ij}\}$ is the annotated segmentation mask, where $l_{ij} \in \{0, \dots, K - 1\}$ is the annotated class label for pixel (i, j) .
- $\mathbf{s} = \{s_{ij}\}$ is an numeric array of per-pixel scores, where s_{ij} is a fine-grained label quality score for pixel (i, j) . Here we consider fine-grained scores computed via some simple function $s_{ij} = f(l_{ij}, p_{ij1}, \dots, p_{ijK})$ applied independently to the information at each pixel (i, j) . Any label quality score for classification data could be used here (Kuan & Mueller, 2022). We stick with the most straightforward choice: $s_{ij} = p_{ijk^*}$ where $k^* = l_{ij}$, i.e. the model-estimated likelihood of the annotated class (Muller & Markert, 2019).

- $\mathbf{b} = \{b_{ij}\}$ is an array of per-pixel binary values, where $b_{ij} = 0$ if the pixel at position (i, j) is flagged as potentially mislabeled (otherwise $b_{ij} = 1$) by a *flagger* LED algorithm (Klie et al., 2022). Here we use Confident Learning (Northcutt et al., 2021b). Such methods can be applied to estimate *which* instances are mislabeled in a classification dataset; here we can simply treat each pixel as a separate instance for a direct application.

Having computed (some of) the above per-pixel information, we are interested in obtaining an overall label quality score s for each image x , to help us prioritize which images require review. To avoid wasted reviewing effort, our primary desiderata is that images with the *lowest* scores should contain some form of annotation error, and images where every pixel is labeled correctly should receive much *higher* scores. An effective score s should be robust to minor statistical fluctuations in per-pixel model outputs which are inevitable in practical deep learning. In this paper, we consider the following 8 methods for producing the label quality score $s(x)$ of an image x . We first present some baseline approaches that help better motivate our proposed *Softmin* method.

3.1. Correctly Classified Pixels (CCP)

This approach measures the proportion of correctly classified pixels in an image. It is calculated as follows:

$$s_{CCP}(x) = \frac{\sum_{i,j} \mathbb{I}[l_{ij} = P_{ij}]}{h \cdot w} \quad (1)$$

Relying on *Correctly Classified Pixels* is an intuitive way to use machine learning to verify the choices made by data annotators. This approach should be robust to small statistical fluctuations in the predicted probability values output by a model, given it only operates on the predicted segmentation mask. However it ignores both model confidence and shortcomings of the trained model.

3.2. Thresholded CCP (TCCP)

The Correctly Classified Pixels method is sensitive to the model’s propensity to over/under predict certain classes relative to others. This alternative approach extends Correctly Classified Pixels with an adaptive threshold parameter and calculating accuracy using a different threshold for each class. We select the threshold value for each class that maximizes accuracy for predicting this class, and use the mean over all classes as our overall label quality score.

For each class $k \in \{1, \dots, K\}$ and threshold τ in a predefined set of thresholds \mathcal{T} , we calculate accuracy as follows:

$$s_{TCCP,t}^k(x) = \frac{\sum_{i,j} \mathbb{I}[l_{ij} = k, p_{ijk} > \tau]}{h \cdot w} \quad (2)$$

For each class k , we select a separate threshold value τ_k^* that maximizes $s_{TCCP,\tau}^k(x)$:

$$\tau_k^* = \operatorname{argmax}_{\tau \in \mathcal{T}} s_{TCCP,\tau}^k(x) \quad (3)$$

The overall TCCP label quality score for an image x is then the mean of these per-class values over all classes:

$$s_{TCCP}(x) = \frac{1}{K} \sum_{k=1}^K s_{TCCP,\tau_k^*}^k(x) \quad (4)$$

Thresholding the probabilistic predictions separately for each class allows us to account for shortcomings of the model if it systematically over/under-predicts a certain class. This can be viewed as a form of post-hoc calibration applied before computing the overall label quality score.

3.3. Confidence In Label (CIL)

We compute the confidence of a given label using the model’s predicted probabilities – simply adopting the model-estimated likelihood of the annotated class at each pixel as a label quality score. Recall that $p_{i,j,c}$ denotes the predicted probability that pixel at position (i, j) belongs to class c , and consider the shorthand: $s_{ij} := p_{i,j,l_{ij}}$, for expressing the model-estimated likelihood of the annotated class at each pixel l_{ij} . Each s_{ij} serves as a quality score for an individual pixel annotation, and the CIL score of each image x is simply the mean of these pixel-scores within the image:

$$s_{CIL}(x) = \frac{1}{h \cdot w} \sum_{i,j} s_{ij} \quad (5)$$

3.4. Softmin (Our Proposed Method)

With a well-trained model, we expect the resulting per-pixel scores s_{ij} to be lower for pixels that are incorrectly annotated or otherwise ambiguous looking. Even in perfectly labeled regions of an image, the s_{ij} will inevitably vary due to natural statistical fluctuations (estimation error in model training). Thus their mean, $s_{CIL}(x)$, may be undesirably sensitive to such s_{ij} variations in correctly-labeled regions of an image (which presumably represent most of most images). This can be mitigated by focusing solely on the *worst-labeled* region of an image, for instance the minimum of the s_{ij} corresponds to one of the least confidently well-labeled pixels in the image. While robust to nuisance variation in the pixel-scores for correctly-labeled regions of the image, the minimum score entirely ignores most of the image, which is also undesirable. Presumably it is easier to determine certain images are mislabeled when a large region of many pixels all receive low scores (see Figure 2).

Instead of taking the minimum of these scores, we can instead form a soft approximation of the minimum to remain robust to nuisance variation in high pixel-scores while

still accounting for the scores across the entire image. Our *softmin* scoring method interpolates between the mean and minimum score. It computes an overall label quality score for the image from per-pixel confidence values via a soft version of the minimum operator:

$$s_{SM}(x) = \sum_{i,j} s_{ij} \cdot \frac{\exp\left(\frac{1-s_{ij}}{\tau}\right)}{\sum_{i,j} \exp\left(\frac{1-s_{ij}}{\tau}\right)} \quad (6)$$

This is analogous to the popular softmax approximation of the argmax operator, where temperature parameter τ controls the sharpness of the minimum approximation above. Since the s_{ij} are confidence values between 0 and 1, we use a smaller $\tau = 0.1$ throughout to appropriately tradeoff between emphasizing the smallest per-pixel scores (most indicative of annotation error locations within an image) while still accounting for the rest of the image. A similar setting was explored by Wang & Mueller (2022) to diagnose mislabeling of text datasets used for entity recognition.

3.5. Confident Learning Counts (CLC)

The alternative *Confident Learning Counts* label quality score extends the popular Confident Learning method for LED in classification (Northcutt et al., 2021b) to segmentation settings. Treating each pixel as an independent instance in a classification task (ignoring which image it belongs to), we apply Confident Learning to infer a binary mask \mathbf{b} estimating *which* pixels in image x are mislabeled. A resulting label quality score is naturally defined as the proportion of pixels in the image estimated to be mislabeled:

$$s_{CLC} = \frac{\sum_{i,j} b_{ij}}{h \cdot w} \quad (7)$$

To reduce runtime and memory requirements of the method, we first downsample predicted/annotated segmentation masks 4 times before computing s_{CLC} . In downsampling, predicted class probabilities are mean pooled within each 4×4 grid, and the resulting annotated label for this grid is determined via majority vote over the 4×4 pixels.

3.6. IOU

A standard evaluation measure to measure the similarity between segmentation predictions and labels is the IoU metric (Intersection over Union, a.k.a. Jaccard Index) (Kirillov et al., 2023; Chan et al., 2021). The IOU for an individual image can serve as a label quality score, calculated as:

$$s_{IOU}(x) = \frac{|\mathbf{P} \cap \mathbf{I}|}{|\mathbf{P} \cup \mathbf{I}|} \quad (8)$$

where \mathbf{P} represents the predicted segmentation mask and \mathbf{I} represents the annotated mask. Sorting images based on

s_{IOU} corresponds to traditional *error analysis*, in which one inspects the highest loss images for suspicious anomalies, with loss computed with respect to model predictions based on a standard evaluation metric.

3.7. Connected Components (CoCo)

Rottmann & Reese (2022) suggest that LED on the pixel-level is less robust than scoring label quality at a connected component level. This approach exploits the spatial nature of segmentation data, where neighboring pixels often belong to the same class. In the *Connected Components* approach, we produce a quality score for each component of an image, a spatially-contiguous region of pixels all annotated/predicted to share the same class. Subsequently we pool these component scores into an overall image score by taking their average over components and classes. Component scores are produced by producing a predicted class probability for the overall component and using it to evaluate the likelihood of the annotated label for the component.

More specifically, we calculate an overall label quality score $s_{CoCo}(x)$ for image x as follows:

1. Form a set of connected components based on each unique combination of predicted and annotated segmentation masks \mathbf{P} and \mathbf{I} .
2. For each connected component c , use the mean of the predicted probabilities p_{ijk} for the pixels within the component as a predicted class probability estimate for the entire component p_c .
3. Compute a quality score for the component $s_c = p_c[k]$ where k is the annotated class label for this component, using the same sort of label-likelihood we computed for each pixel $s_{i,j}$ in other approaches.
4. Average the per-component scores s_c across all components c to obtain the overall label quality score $s_{CoCo}(x)$ for image x .

To reduce runtime and memory requirements of the method, we first downsample predicted/annotated segmentation masks 4 times before computing s_{CoCo} . Here we follow the same downsampling procedure previously described.

4. Experiments

4.1. Datasets

Most real-world segmentation data are full of annotation errors (Vădineanu et al., 2022), but we must benchmark our label quality scores on a dataset where we can be sure of the underlying true segmentation masks. Here we use the SYNTHIA dataset (Ros et al., 2016), a simulated vehicular

dataset similar to Cityscapes (Cordts et al., 2016), but with images generated via graphics engine. To study mislabeling that reflects common issues encountered in real-world segmentation data, we introduce three types of errors into the given labels of our dataset (depicted in Figure 1):

1. **Drop:** Randomly eliminate the label of a selected class, mapping those pixels to the “unlabeled” category. The *Drop* error mimics situations in which annotators forget to label portions of an image, overlooking certain objects, or forgetting the class exists in the set of choices.
2. **Swap:** Randomly interchange the labels of two selected classes across all pixels of the chosen image. The *Swap* error mimics situations where annotators possess low certainty about which class to select or accidentally select the wrong one, perhaps not realizing a better class label exists in the set of categories.
3. **Shift:** Inspired by Heller et al. (2018), this error introduces variability in the shape of a mask for a chosen class, specifically along its edges. Using the OpenCV library (OpenCV, 2023), we introduce random morphological variation in the segmentation masks, representing annotators who sloppily draw the labels.

Our benchmark has three datasets, each containing one of these types of labeling errors. The set of images is the same in each dataset, but differing between datasets is: which images are mislabeled, the types of label errors, and the proportion of mislabeled images. In our first, second, and third dataset: 20%, 30%, and 20% of the images respectively have a *Drop*, *Swap*, or *Shift* error. These settings allow us to characterize label error detection performance separately across the three error types, facilitating fine-grained study of our proposed methods under diverse conditions observed in real-world segmentation data (Vădineanu et al., 2022).

In each dataset, label errors are randomly introduced before establishing training and validation splits. Each training and each validation dataset has 1,112 images. In our benchmark, we only produce predictions and label quality scores for images in the validation set – these are the only images considered in our evaluations. Note that one could instead employ cross-validation to score label quality of every image, but the benchmark conclusions would unlikely change.

4.2. Models

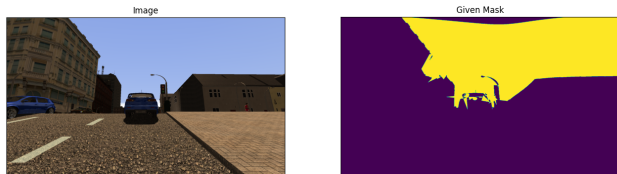
To study how well LED methods generalize across different segmentation models, we apply each label quality scoring technique with two different models:

1. DeepLabV3+ (Chen et al., 2018): enhances the DeepLabV3 model with a decoder module to further refine segmentation results, particularly along object

Drop (Car overlooked) - Given Mask for Unlabeled class



Shift Error (in Sky class) - Given Mask for Sky class



Swap (Building → Vegetation) - Given Mask for Vegetation



Figure 1. Examples of the three types of annotation errors in our SYNTHIA datasets. In the depicted *Drop* error, the car was overlooked by annotators and its pixels are left as part of the Unlabeled class. In the depicted *Shift* error, the mask for the Sky class was poorly drawn and incorrectly covers part of the car. In the depicted *Swap* error, the buildings have been mislabeled as the Vegetation class instead of the Building class. All three of these images are amongst the top-10 images with lowest softmin label quality score in the dataset, and thus automatically prioritized for review.

boundaries. This model has shown excellent performance in semantic segmentation tasks and has been widely used in various applications.

2. FPN (Lin et al., 2017) The Feature Pyramid Network is a strong feature extractor for multi-scale object detection. The FPN model handles scale variation by using a top-down pathway and lateral connections to combine low-resolution, semantically strong features with high-resolution, semantically weak features.

Both models are implemented using the PyTorch backbone provided by Iakubovskii (2019) (with the “se_resnext50_32x4d” encoder and “imagenet” pretrained weights). The final activation function for both models is “softmax2d”. All three models are fit to the noisily labeled training data each with their respective error type and used to produce out-of-sample predictions for a noisy validation set over which we evaluate LED performance. Our experiments reflect a common setup in practical segmentation applications, in order to understand how effectively label errors can be detected under this setup.

4.3. Evaluation

Detecting the mislabeled images within a large dataset is an information retrieval task, and thus we evaluate performance with standard precision/recall metrics. In particular, we evaluate how well our label quality scores are able to *rank* truly mislabeled images above those which are correctly labeled via: the Area Under the Receiver Operating Characteristic Curve (AUROC), the Area Under the Precision-Recall Curve (AUPRC), and **Lift @ T**. Lift measures how many times more prevalent labels errors are within the top- T scoring images vs. all images (and is monotonically related to **Precision @ T**). We consider $T = 100$ or setting it equal to the true number of mislabeled images in each dataset.

Evaluation via Lift assesses scores’ precision, while AUROC and AUPRC assess both precision and recall. In settings where true positives are relatively rare (as our mislabeled images are here), Saito & Rehmsmeier (2015) argue AUPRC is more informative than AUROC.

5. Results

Tables 1-4 report results from a comprehensive evaluation of all aforementioned label quality methods using both models on our variants of the SYNTHIA dataset. The results reveal that our *Softmin* approach consistently delivers the most (or second-most in certain settings) effective detection of labeling errors, irrespective of the error type or the specific model in use. Our proposed method consistently demonstrates superior performance compared to existing strategies that rely on variations of the IoU metric (Rottmann & Reese, 2022). While the *Connected Components* approach does fare slightly better at detecting *Shift* errors according to some metrics, *Connected Components* label quality scores do not work well for detecting *Drop* errors. We find that detecting *Swap* errors is significantly easier than other two types of error, with *Shift* errors being hardest to detect. Visually, we also found this to be the case when manually inspecting our SYNTHIA images for errors.

5.1. Annotation errors in the CityScapes Dataset

Finally, we apply our methodology to audit for mislabeling in the popular CityScapes segmentation dataset, here studying its *fine*-annotations (Cordts et al., 2016). We fit a DeepLabV3+ model to this data and use the model’s out-of-sample predictions with our softmin score to assess each image’s label quality. Figure 2 shows top *naturally-occurring* label errors discovered by this approach, and many more of such overlooked errors were discovered. The full results for CityScapes are provided in the previously linked benchmarks GitHub repository.

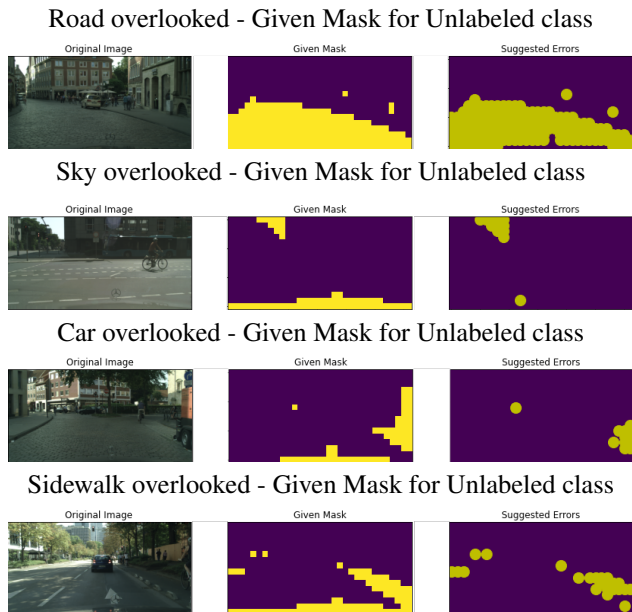


Figure 2. Images which received some of the lowest softmin label quality scores in the Cityscapes dataset. In different images, the *Road*, *Sky*, *Car*, and *Sidewalk* were overlooked by annotators – pixels which should have been annotated with these classes were mistakenly left as the *Unlabeled* class. The Suggested Errors on right highlight the pixels with low s_{ij} values in each image.

6. Discussion

Our study presents a comprehensive evaluation of eight different label quality scoring methods for semantic segmentation datasets. Based on a soft-minimum of the model-estimated likelihoods of each pixel’s annotated class, the *Softmin* score is particularly effective for detecting mislabeled images with high precision and recall, regardless what types of annotation errors lurk in the data. This highlights the value of focusing on potentially severely mislabeled regions of an image when estimating which images have imperfect annotations. Our findings align with insights from Wang & Mueller (2022), who found taking a minimum over per-token scores to be an effective sentence label quality score for entity recognition (text) data.

The *Softmin* score is easy to apply with any trained segmentation model, and thus will remain applicable as improved segmentation architectures and training procedures are invented. After using this method to detect annotation errors, they can be fixed or some of the bad data omitted from the training set, and subsequently a more reliable copy of this same model trained without any change in the modeling code. As the accuracy of models increases via new innovations, the LED performance of our label quality scores will improve, thus enabling *even* more accurate versions of these models to be produced from higher integrity data.

Estimating label quality and errors in semantic segmentation data via any model

Table 1. AUROC achieved by various label quality scores used with two types of models for detecting three types of annotation errors.

METHOD	DROP		SWAP		SHIFT	
MODEL	DEEPLABV3+	FPN	DEEPLABV3+	FPN	DEEPLABV3+	FPN
CORRECTLY CLASSIFIED PIXELS	0.915	0.916	1.000	1.000	0.833	0.798
THRESHOLDED CCP	0.869	0.869	0.993	0.993	0.852	0.818
CONFIDENCE IN LABEL	0.915	0.916	1.000	1.000	0.832	0.797
CONFIDENT LEARNING COUNTS	0.904	0.905	0.999	0.999	0.807	0.778
IOU	0.921	0.901	0.880	0.880	0.713	0.649
CONNECTED COMPONENTS	0.880	0.888	0.984	0.982	0.783	0.819
SOFTMIN	0.951	0.947	0.998	0.998	0.863	0.828

Table 2. AUPRC achieved by various label quality scores used with two types of models for detecting three types of annotation errors.

ERROR TYPE	DROP		SWAP		SHIFT	
MODEL	DEEPLABV3+	FPN	DEEPLABV3+	FPN	DEEPLABV3+	FPN
CORRECTLY CLASSIFIED PIXELS	0.814	0.814	0.999	0.999	0.474	0.399
THRESHOLDED CCP	0.684	0.681	0.983	0.983	0.526	0.440
CONFIDENCE IN LABEL	0.813	0.814	0.999	0.999	0.472	0.398
CONFIDENT LEARNING COUNTS	0.795	0.796	0.999	0.999	0.429	0.375
IOU	0.808	0.749	0.754	0.759	0.440	0.320
CONNECTED COMPONENTS	0.654	0.675	0.971	0.965	0.519	0.537
SOFTMIN	0.888	0.875	0.996	0.996	0.545	0.461

Table 3. Lift @ T achieved by various label quality scores used with two types of models for detecting three types of annotation errors. Here T is set equal to the true number of mislabeled images in each dataset.

ERROR TYPE	DROP		SWAP		SHIFT	
MODEL	DEEPLABV3+	FPN	DEEPLABV3+	FPN	DEEPLABV3+	FPN
CORRECTLY CLASSIFIED PIXELS	3.637	3.662	3.369	3.369	2.391	2.254
THRESHOLDED CCP	3.216	3.117	3.167	3.157	2.709	2.345
CONFIDENCE IN LABEL	3.637	3.662	3.369	3.369	2.413	2.254
CONFIDENT LEARNING COUNTS	3.588	3.588	3.369	3.358	2.163	2.072
IOU	3.983	3.464	2.341	2.320	2.254	1.935
CONNECTED COMPONENTS	3.142	3.167	3.114	3.061	2.550	2.618
SOFTMIN	4.231	4.107	3.305	3.305	2.755	2.550

Table 4. Lift @100 achieved by various label quality scores used with two types of models for detecting three types of annotation errors.

ERROR TYPE	DROP		SWAP		SHIFT	
	DEEPLABV3+	FPN	DEEPLABV3+	FPN	DEEPLABV3+	FPN
CORRECTLY CLASSIFIED PIXELS	5.088	5.088	3.432	3.432	2.315	2.113
THRESHOLDED CCP	4.196	4.144	3.432	3.432	2.767	2.365
CONFIDENCE IN LABEL	5.088	5.088	3.432	3.432	2.315	2.113
CONFIDENT LEARNING COUNTS	4.983	4.983	3.432	3.432	2.164	2.013
IoU	4.773	4.511	3.020	3.055	3.120	2.164
CONNECTED COMPONENTS	3.882	3.934	3.432	3.432	3.170	3.271
SOFTMIN	5.140	5.193	3.432	3.432	3.019	2.315

References

- Chachuła, K., Popowicz, A., Łyskawa, J., Olber, B., Fratzczak, P., and Radlak, K. Combating noisy labels in object detection datasets. *arXiv preprint arXiv:2211.13993*, 2022.
- Chan, R., Rottmann, M., and Gottschalk, H. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation, 2021.
- Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation, 2018.
- Chen, P., Liao, B. B., Chen, G., and Zhang, S. Understanding and utilizing deep neural networks trained with noisy labels. In *International Conference on Machine Learning*, 2019.
- Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. The cityscapes dataset for semantic urban scene understanding. *CoRR*, abs/1604.01685, 2016. URL <http://arxiv.org/abs/1604.01685>.
- Goh, H. W. and Mueller, J. Activelab: Active learning with re-labeling by multiple annotators. In *ICLR Workshop on Trustworthy ML*, 2023.
- Heller, N., Dean, J., and Papanikolopoulos, N. Imperfect segmentation labels: How much do they matter?, 2018.
- Iakubovskii, P. Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch, 2019.
- Jiang, L., Zhou, Z., Leung, T., Li, L.-J., and Fei-Fei, L. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, 2018.
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R. Segment anything, 2023.
- Klie, J.-C., Webber, B., and Gurevych, I. Annotation error detection: Analyzing the past and present for a more coherent future. *arXiv preprint arXiv:2206.02280*, 2022.
- Kuan, J. and Mueller, J. Model-agnostic label quality scoring to detect real-world label errors. In *ICML DataPerf Workshop*, 2022.
- Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. Feature pyramid networks for object detection, 2017.
- Muller, N. M. and Markert, K. Identifying mislabeled instances in classification datasets. In *International Joint Conference on Neural Networks*. IEEE, jul 2019.
- Murrugarra-Llerena, J., Kirsten, L. N., and Jung, C. R. Can we trust bounding box annotations for object detection? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4813–4822, 2022.
- Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. Learning with noisy labels. In *Advances in Neural Information Processing Systems*, 2013.
- Nishi, K., Ding, Y., Rich, A., and Hollerer, T. Augmentation strategies for learning with noisy labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.
- Northcutt, C. G., Athalye, A., and Mueller, J. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of the 35th Conference on Neural Information Processing Systems Track on Datasets and Benchmarks*, December 2021a.
- Northcutt, C. G., Jiang, L., and Chuang, I. L. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021b.
- OpenCV. Opencv (open source computer vision library), 2023. URL <https://opencv.org/>. Open Source Computer Vision Library.
- Ros, G., Sellart, L., Materzynska, J., Vazquez, D., and Lopez, A. M. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3234–3243, 2016. doi: 10.1109/CVPR.2016.352.
- Rottmann, M. and Reese, M. Automated detection of label errors in semantic segmentation datasets via deep learning and uncertainty quantification, 2022.
- Saito, T. and Rehmsmeier, M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLoS one*, 10(3):e0118432, 2015.
- Song, H., Kim, M., Park, D., Shin, Y., and Lee, J.-G. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
- Sukhbaatar, S. and Fergus, R. Learning from noisy labels with deep neural networks. *arXiv preprint arXiv:1406.2080*, 2014.

- Thyagarajan, A., Snorrason, E., Northcutt, C., and Mueller, J. Identifying incorrect annotations in multi-label classification data. *ICLR Workshop on Trustworthy ML*, 2023.
- Vădineanu, Ș., Pelt, D. M., Dzyubachyk, O., and Batenburg, K. J. An analysis of the impact of annotation errors on the accuracy of deep learning for cell segmentation. In *International Conference on Medical Imaging with Deep Learning*, pp. 1251–1267. PMLR, 2022.
- Wang, W.-C. and Mueller, J. Detecting label errors in token classification data. In *NeurIPS Workshop on Interactive Learning for Natural Language Processing*, 2022.
- Wei, J., Zhu, Z., Cheng, H., Liu, T., Niu, G., and Liu, Y. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2022.
- Zhang, L., Tanno, R., Xu, M.-C., Jin, C., Jacob, J., Ciccarelli, O., Barkhof, F., and Alexander, D. C. Disentangling human error from the ground truth in segmentation of medical images. In *Advances in Neural Information Processing Systems*, 2020.
- Zhang, Z. and Sabuncu, M. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems*, 2018.