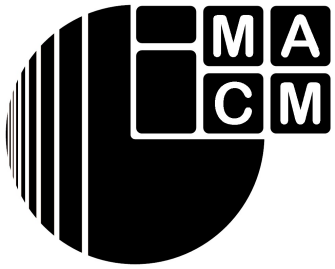

Bergische Universität Wuppertal

Fachbereich Mathematik und Naturwissenschaften

Institute of Mathematical Modelling, Analysis and Computational Mathematics
(IMACM)

Preprint BUW-IMACM 19/06

R. Chan, M. Rottmann, F. Hüger, P. Schlicht and H. Gottschalk

Application of Decision Rules for Handling Class Imbalance in Semantic Segmentation

March 5, 2019

http://www.math.uni-wuppertal.de

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

Application of Decision Rules for Handling Class Imbalance in Semantic Segmentation

Robin Chan **Matthias Rottmann** **Hanno Gottschalk**
School of Mathematics and Natural Sciences
University of Wuppertal
{rchan,rottmann,hgottsch}@uni-wuppertal.de

Peter Schlicht **Fabian Hüger**
Architecture and AI Technologies
Volkswagen Group Research Automated Driving
{peter.schlicht,fabian.hueger}@volkswagen.de

Abstract

As part of autonomous car driving systems, semantic segmentation is an essential component to obtain a full understanding of the car's environment. One difficulty, that occurs while training neural networks for this purpose, is class imbalance of training data. Consequently, a neural network trained on unbalanced data in combination with maximum a-posteriori classification may easily ignore classes that are rare in terms of their frequency in the dataset. However, these classes are often of highest interest. We approach such potential misclassifications by weighting the posterior class probabilities with the prior class probabilities which in our case are the inverse frequencies of the corresponding classes in the training dataset. More precisely, we adopt a localized method by computing the priors pixel-wise such that the impact can be analyzed at pixel level as well. In our experiments, we train one network from scratch using a proprietary dataset containing 20,000 annotated frames of video sequences recorded from street scenes. The evaluation on our test set shows an *increase of average recall* with regard to instances of pedestrians and info signs by 25% and 23.4%, respectively. In addition, we significantly *reduce the non-detection* rate for instances of the same classes by 61% and 38%.

1 Introduction

A common issue with "real world" datasets is the imbalance of observed object classes. Class imbalance in datasets can have a detrimental effect on classification performance of neural networks (NNs) trained on such datsets, see also [4]. Methods overcoming class imbalance can be divided into two main categories [20]. The first category are *sampling-based* methods that operate directly on a dataset with the aim to balance its class distribution. Oversampling and undersampling strategies have been proposed in [4, 14, 20]. In their basic versions, the dataset is balanced by increasing the number of instances from "minority" classes and by decreasing the number of instances from "majority" classes, respectively. A more advanced method called SMOTE [3] combines these two aforementioned approaches, resulting in an additional boost in classification performance.

The second category are *algorithm-based* methods [4, 13, 20]. They make use of cost-based training and decision thresholding. The idea behind these strategies is to assign different costs to classification mistakes for different classes. Accordingly, one possibility is to minimize the misclassification cost instead of the standard loss function [12] during training. This would however bias the softmax probability output of the NN. The other possibility is to make class predictions cost-sensitive during

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

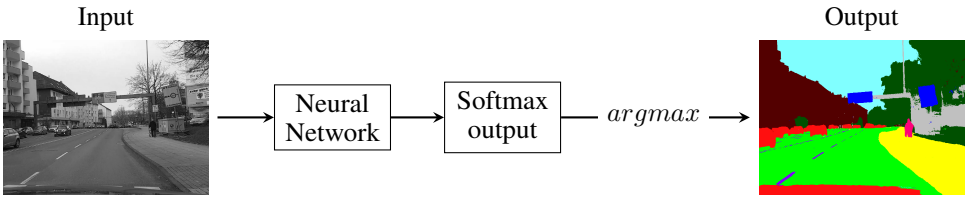

Figure 1: *Illustration of semantic segmentation.*

inference phase after the network is fully trained by moving the output threshold towards inexpensive classes, see [22].

In this work we deal with a proprietary semantic segmentation dataset of the Volkswagen Group and convolutional neural networks (CNNs), more precisely a Full Resolution Residual Network (FRRN) [16], trained on this dataset. In contrast to the publicly available Cityscapes dataset [6], our dataset also contains scenes from highways and country roads. Consequently, classes like humans and traffic signs are underrepresented. Using a CNN trained for this task, the semantic segmentation as a pixel-wise classification is obtained by the maximum a-posteriori probability (MAP), i.e., by applying the $argmax$ function to the pixel-wise softmax output, see figure 1. The CNN as a statistical model aims at minimizing the chance of a misclassification which in decision theory is known as Bayes rule. Another mathematically natural approach from decision theory is the *Maximum Likelihood* (ML) rule. While the MAP / Bayes rule incorporates a prior belief about the semantic classes, the maximum likelihood rule decides only by means of the observed features and chooses the most typical class for the given pattern. From now on, we use the abbreviations *Bayes* and *ML* to refer to these decision rules.

Our approach of using the ML decision rule is motivated by work that studies the influence of risk factors on heart diseases [7, 11]. Given a person's features a decision function is computed to determine whether a patient suffers from a heart disease or not. While the total number of falsely diagnosed patients is increased when using ML instead of Bayes, the number of falsely diagnosed patients, who are actually ill, is significantly reduced. This follows from the substantially rare occurrence of patients with a heart disease that the Bayes rule assumes as a prior belief. The ML rule determines the disease independently of this assumption.

Class balancing is not widely applied to CNNs for semantic segmentation. For instance, class balancing is taken into consideration in Fully Convolutional Networks (FCNs) [19], however the authors decide not to take action since the training data is only moderately unbalanced. Furthermore, in SegNet [2] *median frequency balancing* is used, i.e., the weight assigned each class is the corresponding inverse class frequency multiplied by the median of all class frequencies over the whole dataset. It has been shown empirically that using the computed weights in the loss function during training results in a sharp increase in class average accuracy and a slight increase in mean intersection over union, whereas the global accuracy decreases [2].

With our dataset we pursue the approach to cover a wide variety of everyday street scenes in a preferably unbiased fashion. Thus, we do not want to change the training procedure, i.e., we neither change the loss function used for training nor balance the dataset with respect to the classes. However, we cannot ignore the issue of class imbalance and propose to approach this problem by applying decision rules.

In this work, we analyze the impact of applying the ML instead of the Bayes decision rule for CNNs for semantic segmentation. In other words, the dataset and the training procedure remain unchanged and the decision rules are only interchanged at inference. In contrast to [17] that deals with false positive predictions, our main focus is to reduce false negative predictions. The remainder of this work is structured as follows: In section 2 we explain decision rules in general and in section 3 we employ them in combination with neural networks. Next, we implement ML and evaluate its performance in particular in comparison with Bayes in section 4. For our experiments, we train one FRRN from scratch on our dataset called "DS20k" containing $20,000$ annotated images of traffic scenes in Europe. The DS20k dataset is highly unbalanced, especially with respect to the classes "person" and "info sign" that are significantly underrepresented compared to the remaining classes. This setting is very different from the setting in the *Cityscapes dataset* [6] where these two classes are naturally boosted due to all recorded images showing urban street scenes. Moreover, we apply ML at pixel level in order to handle class imbalance in a position-specific manner.

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

2 Decision rules in discriminant analysis

Discriminant analysis is a multivariate statistical analysis task. Given a population consisting of two or more pre-defined clusters, in which each element of the population belongs to exactly one cluster, one wishes to classify observed data into these distinct groups. Therefore, the objective of discriminant analysis is to find a function that discriminates objects of the population based on observable features, and to predict the object's class affiliation from it.

Let Ω be a population consisting of $N \geq 2$ disjoint subsets. For each element $\omega \in \Omega$ we assume there exists one feature vector $x(\omega) \in S \subset \mathbb{R}^n$. Let

$$X : \Omega \to S$$
$$K : \Omega \to \{1, \ldots, N\}$$

be random variables for feature vector x and class affiliation k, respectively. Then, we define a *decision rule*

$$d : \ S \to \{1, \ldots, N\}$$
$$x(\omega) \mapsto \hat{k}$$

to be a function which assigns an element from the feature space to one class. We say, $d(x) = \hat{k}$ is the predicted class for feature vector x. Furthermore, we describe

(i) the a-priori probability of a random object to belong to class k as

$$p(k) := P(K = k) > 0, \tag{1}$$

(ii) the a-posteriori probability of an object to belong to class k given feature x as

$$p(k|x) := P(K = k|X = x) = \frac{P(K = k, X = x)}{P(X = x)} \tag{2}$$

(iii) and the conditional likelihood of an object of class k having feature x as

$$p(x|k) := P(X = x|K = k) = \frac{P(K = k, X = x)}{P(K = k)} = \frac{P(K = k|X = x)P(X = x)}{P(K = k)}. \tag{3}$$

Usually, none of these probabilities are known and they all need to be estimated. Assuming that this is already accomplished, we define the following decision rules:

The *Bayes* decision rule maps feature vectors x to the class that gives the largest a-posteriori probability. Thus, the decision rule is defined by:

$$d_{Bayes}(x) = \underset{k \in \{1,\ldots,N\}}{\arg\max}\ p(k|x). \tag{4}$$

On the contrary, the *Maximum Likelihood* decision rule maps feature vectors x to the class with the largest conditional likelihood. Thus, the decision rule is defined by:

$$d_{ML}(x) = \underset{k \in \{1,\ldots,N\}}{\arg\max}\ p(x|k) \overset{(3)}{=} \underset{k \in \{1,\ldots,N\}}{\arg\max}\ p(k|x)p(x)/p(k) = \underset{k \in \{1,\ldots,N\}}{\arg\max}\ p(k|x)/p(k). \tag{5}$$

In the latter, the class affiliation k is an unknown parameter that needs to be estimated using the principle of maximum likelihood. The decision rule $d_{ML}(x)$ aims at finding the class k for which the features x are most typical (according to the observed features in the training set), independent of the a-priori probability of the particular class. The difference between these two decision rules lies in the adjustment with the prior class probabilities $p(k)$. Obviously, both decision rules are equal when the prior class distribution is balanced, i.e., $p(1) = \ldots = p(N)$.

3 Decision rules in neural networks for semantic segmentation

Let $x \in \{0, \ldots, 255\}^{m \times n}$ be a (gray-scaled) input image with resolution $m \times n$. In analogy with the previous section, the pixel-wise classification in semantic segmentation is then performed by the

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

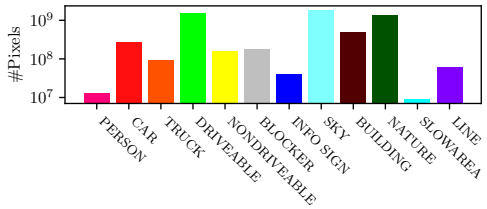

Figure 2: *Illustration of the class imbalance of the DS20k training dataset, for each class we state the number of affiliated pixels in the whole dataset.*

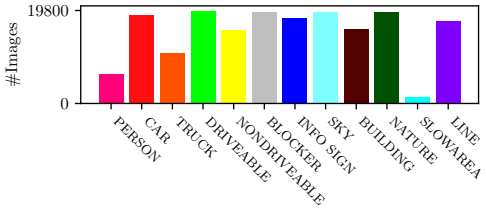

Figure 3: *Illustration of the class imbalance of the DS20k training dataset, for each class we state the number of images in the dataset containing at least one object of the given class.*

$argmax$ function for the estimated class probabilities $p_{ij}(k|x)$ for classes $k \in \{1, \dots N\}$, where $(i, j) \in \{1, \dots, m\} \times \{1, \dots, n\}$ corresponds to the pixel position in the input image and the values for $p_{ij}(k|x)$ are obtained from the softmax output of a segmentation network. This procedure maximizes the overall probability for a correct class estimation which is equivalent to the Bayes rule in decision theory. As stated in [7], this decision rule is optimal for the symmetric cost function

$$c_s\left(\hat{k}, k\right) := \begin{cases} 0 & , \text{ if } \quad \hat{k} = k \\ C & , \text{ if } \quad \hat{k} \neq k \end{cases}, \; C \in \mathbb{R}^+ := (0, \infty) \tag{6}$$

with $\hat{k}$ being the predicted class while k being the target class. This function implies an equal class weighting, also weighting every confusion of two classes, i.e., each type of misclassification, equally. In contrast to that, the ML rule is optimal for the inverse proportional cost function

$$c_p\left(\hat{k}, k\right) := \begin{cases} 0 & , \text{ if } \quad \hat{k} = k \\ C/p(k) & , \text{ if } \quad \hat{k} \neq k \end{cases}, \; C \in \mathbb{R}^+ \tag{7}$$

which increases the cost of a misclassification if the a-priori class probability (in the following called *prior*) is low. Consequently, we need to first determine the class distribution of our dataset in order to apply ML instead of Bayes.

3.1 Class imbalance of data

A statistical analysis of our dataset reveals an unbalanced class distribution in the training set that differs significantly from a uniform distribution, cf. figure 2 and figure 3. For instance, the total number of pixels in *DS20K* belonging to the class PERSON is $1.3 \cdot 10^7$, whereas $1.5 \cdot 10^9$ pixels belong to the class DRIVEABLE. That is a difference in two orders of magnitude. The confusion of these two classes would lead to possibly fatal situations and should be avoided, especially in the domain of near field perception. For the evaluation of the different decision rules, we compute priors and apply the decision rules at pixel level. Note, that the used segmentation network in our experiments will be a fully convolutional neural network which, on an input of infinite size, conserves the translation invariance of convolutional layers. For fully convolutional networks with a receptive field which is small compared to the dimensions of the image, an averaging of pixel-wise priors along the orbits of the translation group (adequately coarsened by pooling) would be adequate [5]. However, the receptive field of our network for a single output pixel contains up to $2/3$ of the input image. Thus, almost all output pixels are affected by boundary effects which enables them to guess their approximate location and justifies our decision for pixel-wise priors. For further discussion we refer to the appendix.

3.2 Computing priors

The priors are essential for the implementation of decision rules. We approximate them using the training set since our network is trained on these unbalanced data. Figure 2 shows the class distribution on full image level in the entire dataset. As there are image regions, where it is more likely that a certain class appears, we are interested in the pixel-wise class distribution of the training dataset. From figure 4 we conclude, that during training there are no pedestrians seen in the upper part of the image. The network thus will be biased towards not predicting a person in that area which might be wrong, e.g., when the street is ascending.

In order to reduce training data specific noise and details from the priors, we smoothen them using a Gaussian filter. Also, a lower cut off limit of 10^{-5} is applied to the priors of all classes, for the purpose of avoiding divisions by zero, see equation (5). A visualization can be found in figure 5.

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

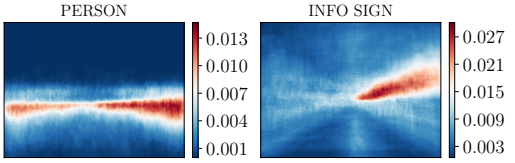

Figure 4: *Pixel-wise priors for class PERSON and INFO SIGN. For determining the priors, the frequency of class appearance at every pixel position in the training set is divided by the total number of training images.*

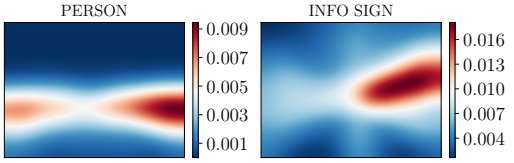

Figure 5: *Smoothed pixel-wise priors for class PERSON and INFO SIGN. For smoothing the priors, a Gaussian-filter is applied with parameter $\sigma = 80$ on the actual priors in order to reduce training data specific noise.*

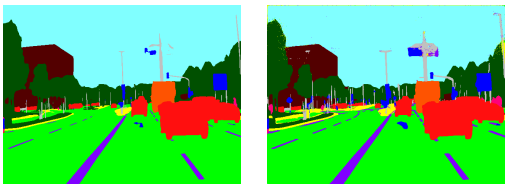

Figure 6: *Visualization of a segmentation with Bayes (Left) and ML (Right). The Bayes prediction serves as benchmark. The ML prediction provides additional insights about the network's segmentation capabilities. Especially the points, where different classes are predicted and thus the decision rules disagree, will be of interest for further analysis.*

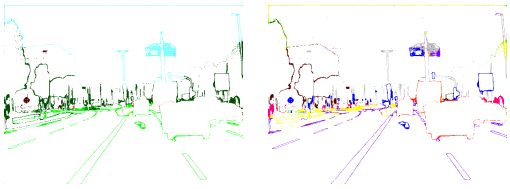

Figure 7: *Visualization of the differences between the segmentations obtained from Bayes and ML. In both images the predictions differ mostly at the object boundaries, i.e., at the transition from one class to another. Left: The points in the Bayes prediction where Bayes and ML disagree. Right: The points in the ML prediction where Bayes and ML disagree.*

4 Numerical experiments with decision rules

We evaluate the performance of Bayes and ML on 200 annotated test images from DS20k that were not used during training. The network we use in our experiments is a slightly modified version of the full-resolution residual network (FRRN) [16] which is a combination of SegNet [2] and ResNet [9] and performs well with respect to recognition and localization. Unlike the original version of the network we use dropout regularization [21] instead of batch normalization. Furthermore, we modify the number of channels per layer.

We implement our FRRN with Tensorflow [1] and train the network by minimizing the negative log-likelihood loss function using the ADAM optimizer [18]. Besides that, we train the network on 19,800 annotated training images (resolution 640×480 pixels) of traffic scenes with 12 different semantic classes. We use a batch size of 16, resulting in a training time of 1 day and 16 hours for 100 epochs with 4 NVIDIA Geforce GTX 1080Tis.

4.1 Visual comparison of Bayes and Maximum Likelihood decision rule

Before we start a quantitative analysis on the impact of ML we visualize the segmentations obtained by both decision rules in order to get a basic understanding of the differences, see figure 6.

At first glance we observe that there are no significant differences between both segmentations. Both decision rules produce the same output for most of the pixels which indicates that the network is well-trained and predicts with high confidence. In particular, the predictions for frequent classes, such as road, nature, sky and buildings, enhance this impression because they make up a large portion of the image. Therefore, the pixel positions, where the decision rules produce different predictions, are of special interest for us.

On closer inspection we observe that the Bayes and ML predictions differ remarkably often at the transition from one class to another. While Bayes prefers to predict the more frequent class at these pixel positions, ML does the opposite and prefers the less frequent class. Thus, ML enlarges the size of "minority" class objects compared to Bayes for the purpose of decreasing the risk of missing any pixels belonging to rare classes. Moreover, we observe that ML produces many (in terms of class affiliation) isolated false positive pixels which are boosted by the priors. For instance, in figure 7 ML frequently classifies scattered pixels as NONDRIVEABLE (pavement, traffic island,...) in the upper part of the image. Since the class $k = \text{NONDRIVEABLE}$ has an extremely small

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

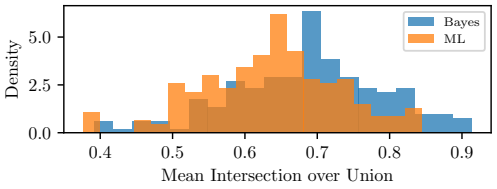

Figure 8: *Histogram of mIoU scores considering all 200 test images*

Decision rule	Overall	PERSON	INFO
Bayes	68.8	40.7	38.7
ML	63.6	22.3	29.5

Table 1: *mIoU scores (in percent) additionally averaged over all images.*

Class	Bayes	ML	GT
PERSON	289	857	368
INFO	686	1592	1349

Table 2: *Number of connected components of class PERSON and INFO.*

	Precision		Recall		IoU	
Class	Bayes	ML	Bayes	ML	Bayes	ML
PERSON	61.1	37.4	48.1	73.1	31.7	27.5
INFO	67.6	36.9	33.6	57.0	23.4	21.2

Table 3: *Average precision, recall and IoU scores of a connected component in prediction*

a-priori probability in the upper part of the image, we see that even small posterior probabilities $p(k|x)$ can result in $p(k|x)/p(k)$ being dominant and thus k being the predicted class when using the ML decision rule. In order to reduce uncertainty, we employ Monte-Carlo dropout, see [8, 10], at inference by computing 10 different predictions under dropout and averaging the output probabilities.

Due to the occurrence of very local misclassifications we add two post processing steps: First, we discard all connected components of one class that contain less than 10 pixels in our statistical computations. Second, we treat connected components of the same class which have less than 10 pixels in-between as one connected component.

4.2 Experiments with Bayes and Maximum Likelihood decision rule

Let $A_{k,\hat{k}} \in \mathbb{N}_0$ be the number of pixels of class k predicted to belong to class $\hat{k}$. For the evaluation, we compute three different performance measures that are common in semantic segmentation:

(i) Precision

$$prc_j = A_{j,j} / \sum_{k=1}^{N} A_{k,j},\ j = 1, \dots, N \tag{8}$$

(ii) Recall

$$rec_j = A_{j,j} / \sum_{k=1}^{N} A_{j,k},\ j = 1, \dots, N \tag{9}$$

(iii) Intersection over Union

$$IoU_j = A_{j,j} / \left(A_{j,j} + \sum_{k=1, k\neq j}^{N} (A_{k,j} + A_{j,k}) \right),\ j = 1, \dots, N. \tag{10}$$

Mean Intersection over Union. We compare the segmentation performance of Bayes and ML, first in terms of *mean intersection over union* (mIoU) which is the average value over all classes for the intersection over union (IoU), i.e., $mIoU = \sum_{k=1}^{N} IoU_k / N$.

Figure 8 shows that Bayes is superior to ML regarding the overall performance. In nearly all images the mIoU for Bayes is higher than for ML. Further averaging of the mIoU over all test images in table 1 reveals a difference of more than 5% in mIoU. This finding is not unexpected since Bayes maximizes the overall probability of correct class predictions. We also compare the IoUs for the rare classes PERSON and INFO SIGN (short hand: INFO). They even show an increased superiority of Bayes. In the subsequent paragraphs we show that this is indeed caused by an overproduction of false positives when using ML.

Precision vs. Recall. Since ML produces many additional false positives for rare classes, we would hope in this case to obtain less false negatives with ML compared to Bayes. By comparing the total number of connected components in the differently predicted segmentations and in the corresponding ground truth segmentation, see table 2, we immediately take notice of a significant impact of ML: the number of connected components in ML segmentations exceeds the number of connected components

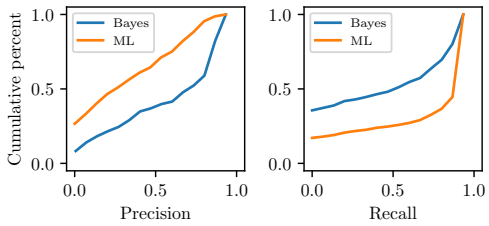

Figure 9: *Cumulative distribution function for precision and recall of class PERSON*

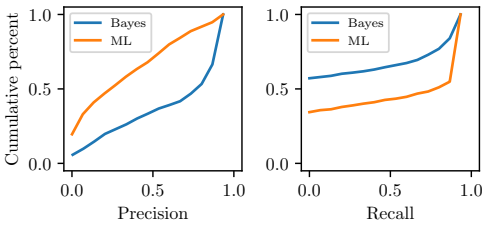

Figure 10: *Cumulative distribution function for precision and recall of class INFO*

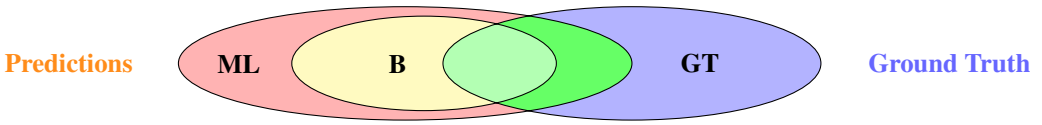

Figure 11: *Graphical illustration of the relation between Bayes and ML prediction segments for rare classes. The ellipse **ML** denotes an segment predicted with ML, analogously **B** denotes the prediction of the same segment with Bayes. Note that $B \subseteq ML$. The ellipse **GT** denotes the corresponding ground truth object. The intersections colored green indicate the true positive pixel predictions, i.e., light green corresponds to the Bayes prediction and dark green to the additional correct pixel predictions obtained with ML. The red color indicates the additional false positive pixel predictions obtained with ML.*

in ground truth (GT) segmentations for both, the PERSON and INFO classes. In contrast to this, Bayes overlooks a significant amount of components.

Consequently, as we expect whole instances to be false positive in ML segmentations but also to recognize more of the rare objects compared to Bayes, we now analyze segment-wise precision and recall in more detail. For this purpose, we define that a (selection of) connected component(s) predicted by some decision rule is a correct object prediction, if there is a ground truth connected component of the same class with non-empty intersection.

The empirical cumulative distribution functions (CDFs) of class PERSON for precision and recall can be found in figure 9. Let F_1 and F_2 be two CDFs, then F_1 is *dominated stochastically to 1st order* by F_2 [15],

$$F_1 \prec F_2, \text{ if } F_1(x) \geq F_2(x) \; \forall \; x. \tag{11}$$

In the following, we denote the CDFs of the Bayes decision rule regarding precision and recall by F_B^p and F_B^r, respectively. Analogously, F_{ML}^p, F_{ML}^r refer to the ML decision rule.

As to be expected, we observe a clear advantage of Bayes in terms of precision since $F_{ML}^p \prec F_B^p$. For any precision value x, in particular for low precision values, the frequency that one instance's precision is below x is significantly less with Bayes than with ML. The average difference is about 25%. Hence, Bayes predicts PERSON segments with better precision than ML.

In terms of recall, we observe the opposite behavior: $F_B^r \prec F_{ML}^r$, i.e., ML is superior over Bayes in this metric. The average difference is about 23%. However, for both decision rules the number of non-detected segments, i.e., $F_B^r(0) = 0.36$ and $F_{ML}^r(0) = 0.17$, is quite high.

Qualitatively, we observe very similar results for the INFO class (as with PERSON), see figure 10. In our studies of precision and recall we also observe that the findings from figure 7 hold statistically, i.e., we observe for rare classes that all predicted Bayes segments lie entirely inside of predicted ML segments. A graphical illustration of this is given in figure 11.

False-detection vs. Non-detection. The benefit from applying ML instead of Bayes lies mainly with the reduction of non-detected ground truth objects. Therefore, it is reasonable to analyze the quantity of the latter, especially in relation to the amount of predicted false positive segments. Additionally, we analyze the false and non-detection frequencies in relation to the size of the predicted segment and the actual ground truth segment, respectively.

Figure 12 depicts the number of false positive PERSON segments depending on the size of the instance in the segmentation for ML and Bayes. In the left panel, there is a noticeable decrease in the frequency of false positive ML segments if the segment size increases, i.e., larger predicted segments

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

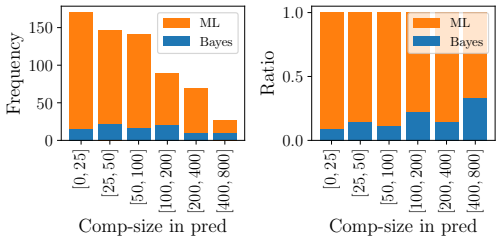

Figure 12: *False-detection of class PERSON conditioned on segment size in prediction*

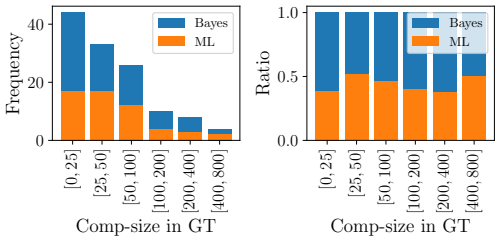

Figure 13: *Non-detection of class PERSON conditioned on object size in ground truth*

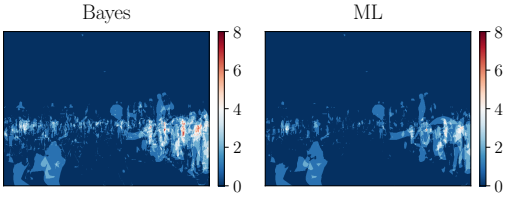

Figure 14: *Heat map of non-detected pixels of class PERSON*

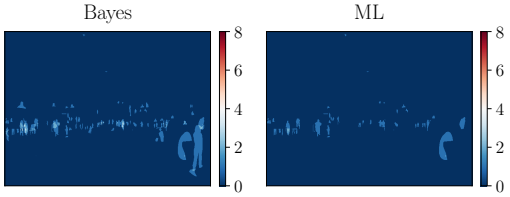

Figure 15: *Heat map of non-detected objects of class PERSON*

are less likely to be entirely incorrect. For Bayes segments, the same tendency holds, even though not as strictly as for ML. Moreover, we see for every segment size bin that the amount of false positive ML segments considerably exceeds the amount of false positive Bayes segments. The right-hand panel of figure 12 shows the amount of Bayes false positives relative to the amount of ML false positives for different component sizes. For increasing component size there is also an increase in the relative amount of Bayes false positives.

Analogous to figure 12, figure 13 shows the number of entirely non-detected objects depending on their size. We observe a similar behavior for Bayes and ML like for the false-detections, also with respect to the object size. For the non-detection of the class PERSON we find a clear advantage in favor of ML, independent of the object size. Bayes overlooks roughly twice as many objects as ML does. This result indicates an uncertainty of the network in finding the rare class PERSON which can be alleviated by using ML. Figure 14 and figure 15 visualize the non-detection at pixel and at object level, respectively.

For the class INFO we observe an analogous behavior. We refer to the appendix for the analysis and corresponding figures.

5 Conclusion and Outlook

In this work, we conducted an in-depth comparison of the ML and Bayes decision rules for a semantic segmentation network trained on an unbalanced dataset showing street scenes. In our tests we observe that ML is able to detect the rare classes PERSON and INFO more frequently than Bayes. Indeed, the pixels that Bayes and ML classify differently indicate that a less frequent class might be overlooked. We have seen that ML detects significantly more instances of rare classes in comparison to Bayes, but to the the detriment of producing even substantially more false-detections which makes ML not reliable for always predicting rare classes correctly. Apart from this, it is important to emphasize that ML only post-processes the softmax output of a neural network. This can be done simultaneously while applying the usual Bayes rule. In the end, we obtain two prediction masks and the additional ML mask is produced computationally nearly for free. What remains is to develop methods to draw plausible conclusions in order to combine both segmentations. Furthermore, the ML prediction can serve as uncertainty mask revealing labeling mistakes of training data or indicating new unlabeled images of high prediction uncertainty which then can be annotated and included in the training process in the manner of active learning. We make the source code of our analysis tool publicly available on GitHub: `https://github.com/robin-chan/decision-rules`.

Acknowledgements. This work is funded in part by Volkswagen Group Research.

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

References

[1] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation". In: *CoRR* abs/1511.00561 (2015). arXiv: 1511.00561. URL: http://arxiv.org/abs/1511.00561.

[3] Kevin W. Bowyer et al. "SMOTE: Synthetic Minority Over-sampling Technique". In: *CoRR* abs/1106.1813 (2011). arXiv: 1106.1813. URL: http://arxiv.org/abs/1106.1813.

[4] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. "A systematic study of the class imbalance problem in convolutional neural networks". In: *CoRR* abs/1710.05381 (2017). arXiv: 1710.05381. URL: http://arxiv.org/abs/1710.05381.

[5] Taco S. Cohen and Max Welling. "Group Equivariant Convolutional Networks". In: *CoRR* abs/1602.07576 (2016). arXiv: 1602.07576. URL: http://arxiv.org/abs/1602.07576.

[6] Marius Cordts et al. "The Cityscapes Dataset for Semantic Urban Scene Understanding". In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[7] L. Fahrmeir, A. Hamerle, and W. Häussler. *Multivariate statistische Verfahren*. 2nd ed. Walter De Gruyter, 1996. ISBN: 978-3110138061.

[8] Yarin Gal and Zoubin Ghahramani. "Dropout As a Bayesian Approximation: Representing Model Uncertainty in Deep Learning". In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*. ICML'16. New York, NY, USA: JMLR.org, 2016, pp. 1050–1059. URL: http://dl.acm.org/citation.cfm?id=3045390.3045502.

[9] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *CoRR* abs/1512.03385 (2015). arXiv: 1512.03385. URL: http://arxiv.org/abs/1512.03385.

[10] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. "Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding". In: *CoRR* abs/1511.02680 (2015). arXiv: 1511.02680.

[11] D.G. Kleinbaum and Lawrence L. Kupper. *Applied Regression Analysis and Other Multivariable Methods*. Duxbury Press, 1978. ISBN: 9780871503558.

[12] Matjaz Kukar and Igor Kononenko. "Cost-Sensitive Learning with Neural Networks". In: (1998), pp. 445–449.

[13] Victoria López et al. "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics". English. In: *Information Sciences* 250 (Nov. 2013), pp. 113–141. ISSN: 0020-0255. DOI: 10.1016/j.ins.2013.07.007.

[14] Ajinkya More. "Survey of resampling techniques for improving classification performance in unbalanced datasets". In: (2016). arXiv: 1608.06048.

[15] G. Ch. Pflug and W. Römisch. *Modeling Maeasuring and Managing Risk*. 1st ed. World Scientific, 2007. ISBN: 978-9812707406.

[16] T. Pohlen et al. "Full-Resolution Residual Networks for Semantic Segmentation in Street Scenes". In: *ArXiv e-prints* (Nov. 2016). arXiv: 1611.08323 [cs.CV].

[17] Matthias Rottmann et al. "Prediction Error Meta Classification in Semantic Segmentation: Detection via Aggregated Dispersion Measures of Softmax Probabilities". In: *CoRR* abs/1811.00648 (2018). arXiv: 1811.00648. URL: http://arxiv.org/abs/1811.00648.

[18] Sebastian Ruder. "An overview of gradient descent optimization algorithms". In: *CoRR* abs/1609.04747 (2016). arXiv: 1609.04747. URL: http://arxiv.org/abs/1609.04747.

[19] Evan Shelhamer, Jonathan Long, and Trevor Darrell. "Fully Convolutional Networks for Semantic Segmentation". In: *PAMI* (2016). URL: http://arxiv.org/abs/1605.06211.

[20] H. Small and J. Ventura. "Handling Unbalanced Data in Deep Image Segmentation". In: (2017). URL: http://cs.uccs.edu/~jkalita/work/reu/REU2017/16Small.pdf.

[21] Nitish Srivastava et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15 (2014), pp. 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

[22] Zhi-Hua Zhou and Xu-Ying Liu. "Training cost-sensitive neural networks with methods addressing the class imbalance problem". In: *IEEE Transactions on Knowledge and Data Engineering* 18.1 (2006), pp. 63–77. ISSN: 1041-4347. DOI: 10.1109/TKDE.2006.17.

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

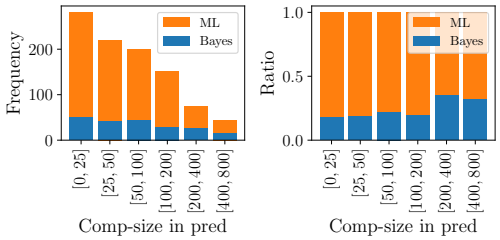

Figure 16: *False-detection of class INFO conditioned on instance size in prediction*

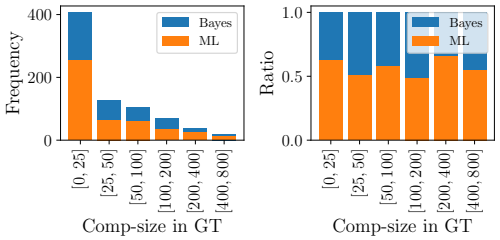

Figure 17: *Non-detection of class INFO conditioned on object size in ground truth*

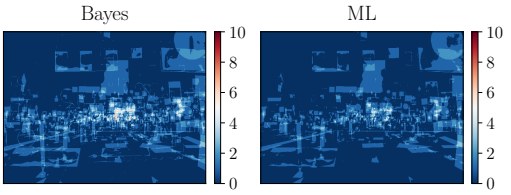

Figure 18: *Heat map of non-detected pixels of class INFO*

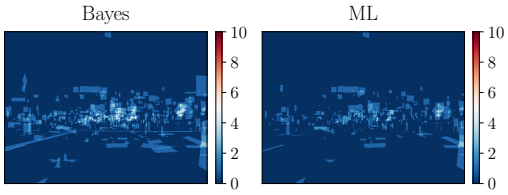

Figure 19: *Heat map of non-detected objects of class INFO*

Appendix

False-detection vs. Non-detection for INFO. Investigating the amount of false-detections as shown in the left panel of figure 16, we notice a staircase-like decrease of false positive ML and Bayes segments for increasing segment size. Moreover, in general there are also more false-detections with ML than with Bayes which can be studied comprehensively in terms of the ratio of false-detections, see right panel of figure 16. On average the amount of Bayes false positives is about 23% of the amount of ML false positives. A slight increase of this percentage can be observed for larger component sizes.

The left panel of figure 17 shows that less info signs are entirely overlooked with ML than with Bayes. For both rules the frequency decreases for larger objects. Nevertheless, most objects of the considered class are rather small which noticeably affects the frequency of non-detections. From the ratio in the right panel of figure 17, we conclude that Bayes fails to detect approximately every second info sign, while ML fails to detect the same every third time. Compared to the performance for class PERSON, ML does not outperform Bayes on detecting info signs as greatly as on detecting persons, but still provides a remarkable performance gain. A visualization of the just mentioned benefit of ML is given in figure 18 and figure 19.

Local Priors vs. Global Prior. In section 3, we introduced a method that uses pixel-wise "local" priors in order to handle the class imbalance depending on the location in the image. However, one might argue that a fully convolutional segmentation network is translation invariant (up to some stride caused by pooling) and the choice of priors should take this into account. We justify our preference over non-localized "global" priors with the network's large receptive field. To this end, we illustrate the impact regarding the non-detection of a PERSON instance when using local priors in comparison to a global prior, see table 4, in particular when the local priors are lower than the global prior.

For class $k = 1, \dots, N$, let $p^g(k)$ be the global prior and let $p^l_{ij}(k)$ be the local prior at pixel position $(i, j) \in \{1, \dots, m\} \times \{1, \dots, n\}$. Then, we denote by

$$B_k := \{(i, j) : p^g(k) \leq p^l_{ij}(k)\} \tag{12}$$

$$B'_k := \{(i, j) : p^g(k) > p^l_{ij}(k)\} \tag{13}$$

the set of pixel positions where the global prior is lower than or equal to the local prior and the set of pixel positions where the global prior is higher than the local prior, respectively.

For our test, we place one PERSON instance at $(i, j) \in B'_{PERSON}$, see figure 20 left image. Since, in this region of the image the network rather expects to see an info sign and the local priors of class PERSON and INFO are of a similar magnitude, we cropped the image such that the PERSON

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

Class	Global Prior
PERSON	0.0022
INFO SIGN	0.0067

Table 4: *Global prior of class PERSON and INFO which is the average value over all pixel-wise priors of the respective class, i.e., the proportion of all pixels in the training set belonging to that class (see also figure 2).*

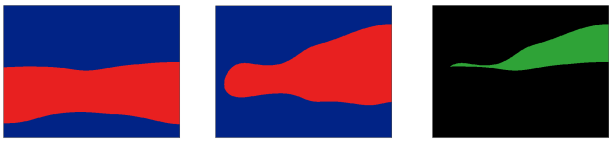

Figure 20: *Visualization of the two sets B_k (red) and B'_k (blue) for the classes PERSON (left panel) and INFO (center panel). In addition, the green color (right panel) shows $B'_{PERSON} \cap B_{INFO}$, i.e., the region of pixel positions in which the local priors are lower than the global prior for class the PERSON and the other way round for INFO.*

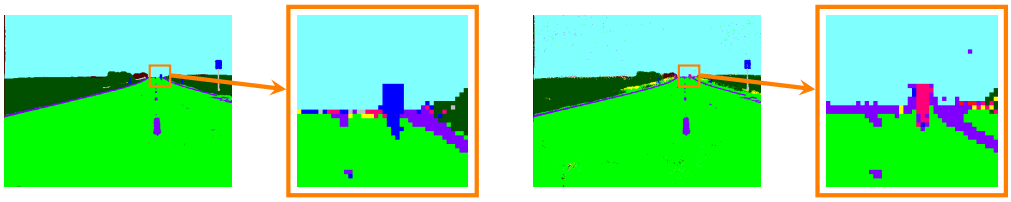

Figure 21: *Example of a non-detected person by using a global prior (two left hand panels) which is detected by using local priors (two right hand panels).*

instance is located at $(i, j) \in B'_{PERSON} \cap B_{INFO}$ in order to provoke a misclassification by using global priors, see figure 20 center and right image.

Figure 21 shows the segmentations produced by using the different priors. We observe that the person is entirely overlooked and predicted to be an info sign by using the global prior while the person is nearly fully detected by using the local priors. Although, this situation is artificially created, it illustrates the importance of using priors and, in particular, the positive effect of localized priors for images outside of the network's learned experience.