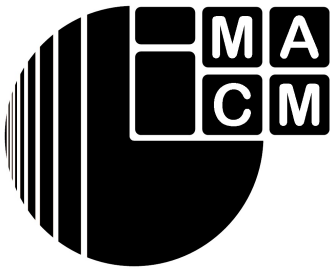


Bergische Universität Wuppertal

Fachbereich Mathematik und Naturwissenschaften

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

Preprint BUW-IMACM 18/18

Matthias Rottmann, Pascal Colling, Thomas Paul Hack, Fabian Hüger, Peter Schlicht and Hanno Gottschalk

Prediction Error Meta Classification in Semantic Segmentation: Detection via Aggregated Dispersion Measures of Softmax Probabilities

December 21, 2018

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

Prediction Error Meta Classification in Semantic Segmentation: Detection via Aggregated Dispersion Measures of Softmax Probabilities

Matthias Rottmann[*], **Pascal Colling**[*], **Thomas Paul Hack**[†], **Fabian Hüger**[‡], **Peter Schlicht**[‡] and **Hanno Gottschalk**[*]

Abstract

We present a method that "meta" classifies whether segments predicted by a semantic segmentation neural network intersect with the ground truth. To this end, we employ measures of dispersion for predicted pixel-wise class probability distributions, like classification entropy, that yield heat-maps of the input scene's size. We aggregate these dispersion measures segment-wise and derive metrics that are well-correlated with the segment-wise IoU of prediction and ground truth. In our tests, we use two publicly available DeepLabv3+ networks (pre-trained on the Cityscapes data set) and analyze the predictive power of different metrics and different sets of metrics. To this avail, we compute logistic LASSO regression fits for the task of classifying $IoU = 0$ vs. $IoU > 0$ per segment and obtain classification rates of up to 81.91% and AUROC values of up to 87.71% without the incorporation of advanced techniques like Monte-Carlo dropout. We complement these tests with linear regression fits to predict the segment-wise IoU and obtain prediction standard deviations of down to 0.130 as well as R^2 values of up to 81.48%. We show that these results clearly outperform standard single-metric baseline approaches.

1 Introduction

In recent years, deep learning has outperformed other classes of predictive models in many applications. In some of these, e.g. autonomous driving or diagnostics in medicine, the reliability of a prediction is of highest interest. In classification tasks, the thresholding on the highest softmax probability or thresholding on the entropy of the classification distributions (softmax output) are commonly used approaches to detect false predictions of neural networks, see e.g. [8, 13]. Metrics like classification entropy or the highest softmax probability are usually combined with model uncertainty (Monte-Carlo (MC) dropout inference) and sometimes input uncertainty, cf. [7] and [13], respectively. These approaches have proven to be practically efficient for detecting uncertainty. Such methods have also been transferred to semantic segmentation tasks. See [15] for further uncertainty metrics proposed recently. The work presented in [12] makes use of MC dropout to model the uncertainty of segmentation networks and also shows performance improvements in terms of segmentation accuracy. This approach was applied in other works to model the uncertainty and filter out predictions with low reliability, cf. e.g. [11, 19]. In [9] this line of research was further developed to detect spacial and temporal uncertainty in the semantic segmentation of videos.

In this work we establish an approach for efficiently meta-classifying whether an inferred segment (representing a predicted object) of a semantic segmentation intersects with the ground truth or not, as similarly proposed for classification problems in [8]. The term *meta classification* has been used in

[*]University of Wuppertal, School of Mathematics and Natural Sciences, rottmann@math.uni-wuppertal.de, {pascal.colling,hanno.gottschalk}@uni-wuppertal.de

[†]University Leipzig, Faculty of Physics, thomas-paul.hack@itp.uni-leipzig.de

[‡]Volkswagen Research, {fabian.hueger,peter.schlicht}@volkswagen.de

Preprint. Work in progress.

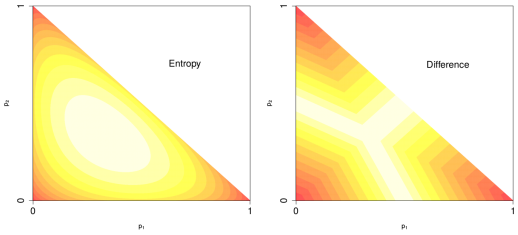


Figure 1: *Visualization of entropy (cf. (2), left panel) and difference (cf. (3), right hand panel) for three variables p_1, p_2 and p_3 treating p_3 implicitly via $p_1 + p_2 + p_3 = 1$.*

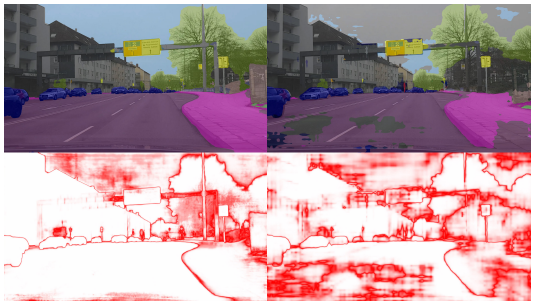


Figure 2: *Example of segmentation (top line) and heat map D_z (bottom line) for Xception65 (left column) and MobilenetV2 (right hand column). Original image is not part of the Cityscapes dataset.*

the context of classical machine learning for learning the weights for each member of a committee of classifiers [14]. In terms of deep learning we use it as a shorthand to distinguish between a network's own classification and the classification whether a prediction is "true" or "false". In contrast to the work cited above, we aim at judging the statistical reliability of each segment inferred by the neural network. To this end, dispersion measures, like entropy, are applied to the softmax probabilities (the networks output) on pixel level yielding dispersion heat maps. We aggregate these heat maps over predicted segments alongside with other quantities derived from the network's prediction like the segment's size and predicted class. From this, we construct per-segment metrics. A commonly used performance measure for the quality of a segmentation is the intersection over union (IoU a.k.a. Jaccard index [10]) of prediction and ground truth. We use the constructed metrics as inputs to logistic regression models for meta classifying, whether an inferred segment's IoU vanishes or not, i.e., predicting $IoU = 0$ or $IoU > 0$. Also, we use linear regression models for predicting a segment's IoU directly, thus obtaining statements about the reliability of the network's prediction.

In our tests, we employ two different publicly available DeepLabv3+ networks [2] that were trained on the Cityscapes dataset [4]. We perform all tests on the Cityscapes validation set and demonstrate that our segment-wise metrics are well correlated with the IoU; thus they are suitable for detecting false positives on segment level. For logistic regression fits we obtain values of up to 87.71% for the area under curve corresponding to the receiver operator characteristic curve (AUROC, see [6]). Predicting the segment-wise IoU via linear regression we obtain prediction standard deviations of down to 0.130 and R^2 values of up to 81.48%.

2 Pixel-wise dispersion metrics and aggregation over segments

A segmentation network with a softmax output layer can be seen as a statistical model that provides for each pixel z of the image a probability distribution $f_z(y|x,w)$ on the q class labels $y \in \mathcal{C} = \{y_1, \ldots, y_q\}$, given the weights w and the data x. The predicted class in z is then given by

$$\hat{y}_z(x,w) = \arg\max_{y \in \mathcal{C}} f_z(y|x,w). \tag{1}$$

Dispersion or concentration measures quantify the degree of randomness in $f_z(y|x,w)$. Here, we consider two of those measures: *entropy* E_z (also known as *Shannon information* [17]) and *difference in probability* D_z, i.e. the difference between the two largest softmax values:

$$E_z(x,w) = -\frac{1}{\log(q)} \sum_{y \in \mathcal{C}} f_z(y|x,w) \log f_z(y|x,w)\,, \tag{2}$$

$$D_z(x,w) = 1 - f_z(\hat{y}_z(x,w)|x,w) + \max_{y \in \mathcal{C} \setminus \{\hat{y}_z(x,w)\}} f_z(y|x,w)\,. \tag{3}$$

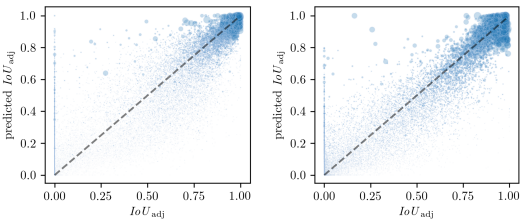


Figure 3: *IoU_{adj} vs. predicted IoU_{adj} for all connected components predicted by Xception65 (left) and MobilenetV2 (right). Dot sizes are proportional to S.*

	XC	MN		XC	MN
$\bar{E}$	-0.70139	-0.70162	$\bar{D}$	-0.85211	-0.84858
$\bar{E}_{bd}$	-0.44065	-0.41845	$\bar{D}_{bd}$	-0.60308	-0.52163
$\bar{E}_{in}$	-0.71623	-0.69884	$\bar{D}_{in}$	-0.85458	-0.82171
$\tilde{\bar{E}}$	0.31219	0.36261	$\tilde{\bar{D}}$	0.22797	0.30245
$\tilde{\bar{E}}_{in}$	0.39195	0.42806	$\tilde{\bar{D}}_{in}$	0.29279	0.35131
S	0.30442	0.47978	$\tilde{S}$	0.50758	0.71071
S_{bd}	0.44625	0.62713	$\tilde{S}_{in}$	0.50758	0.71071
S_{in}	0.30201	0.47708			

Table 1: *Correlation coefficients ρ with respect to IoU_{adj}. Results are computed on the Cityscapes validation set, XC: DeepLabv3+Xception65 and MN: DeepLabv3+MobilenetV2.*

For better comparison, both quantities have been written as dispersion measures and been normalized to the interval $[0, 1]$: One has $E_z = D_z = 1$ for the equiprobability distribution $f_z(y|x, w) = \frac{1}{q}$, $y \in \mathcal{C}$, and $E_z = D_z = 0$ on the deterministic probability distribution ($f_z(y|x, w) = 1$ for one class and 0 otherwise). For the discussion of further dispersion measures, cf. [5]. Figure 1 displays these quantities for three class probability distributions. The most direct method of uncertainty quantification on an image is the heat mapping of a dispersion measure as in fig. 2.

For a given image x we denote by $\hat{\mathcal{K}}_x$ the set of connected components (segments) in the predicted segmentation $\hat{\mathcal{S}}_x = \{\hat{y}_z(x, w)|z \in x\}$ (omitting the dependence on the weights w). Analogously we denote by $\mathcal{K}_x$ the set of connected components in the ground truth $\mathcal{S}_x$. For each $k \in \hat{\mathcal{K}}_x$, we define the following quantities:

- the interior $k_{in} \subset k$ where a pixel z is an element of k_{in} if all eight neighbouring pixels are an element of k
- the boundary $k_{bd} = k \setminus k_{in}$
- the intersection over union IoU: let $\mathcal{K}_x|_k$ be the set of all $k' \in \mathcal{K}_x$ that have non-trivial intersection with k and whose class label equals the predicted class for k, then

$$IoU(k) = \frac{|k \cap K'|}{|k \cup K'|}, \qquad K' = \bigcup_{k' \in \mathcal{K}_x|_k} k'$$

- adjusted IoU_{adj}: let $Q = \{q \in \hat{\mathcal{K}}_x : q \cap K' \neq \emptyset\}$, for reasons explained in the appendix we use in our tests

$$IoU_{\text{adj}}(k) = \frac{|k \cap K'|}{|k \cup (K' \setminus Q)|}$$

- the pixel sizes $S = |k|$, $S_{in} = |k_{in}|$, $S_{bd} = |k_{bd}|$
- the mean entropies $\bar{E}$, $\bar{E}_{in}$, $\bar{E}_{bd}$ defined as

$$\bar{E}_\sharp(k) = \frac{1}{S_\sharp} \sum_{z \in k_\sharp} E_z(x), \qquad \sharp \in \{\_, in, bd\}$$

- the mean distances $\bar{D}$, $\bar{D}_{in}$, $\bar{D}_{bd}$ defined in analogy to the mean entropies
- the relative sizes $\tilde{S} = S/S_{bd}$, $\tilde{S}_{in} = S_{in}/S_{bd}$
- the relative mean entropies $\tilde{\bar{E}} = \bar{E}\tilde{S}$, $\tilde{\bar{E}}_{in} = \bar{E}_{in}\tilde{S}_{in}$, and relative mean distances $\tilde{\bar{D}} = \bar{D}\tilde{S}$, $\tilde{\bar{D}}_{in} = \bar{D}_{in}\tilde{S}_{in}$

Typically, E_z and D_z are large for $z \in k_{bd}$. This motivates the separate treatment of interior and boundary measures. With the exception of IoU and IoU_{adj}, all scalar quantities defined above can be computed without the knowledge of the ground truth. Our aim is to analyze to which extent they are able to predict IoU_{adj}.

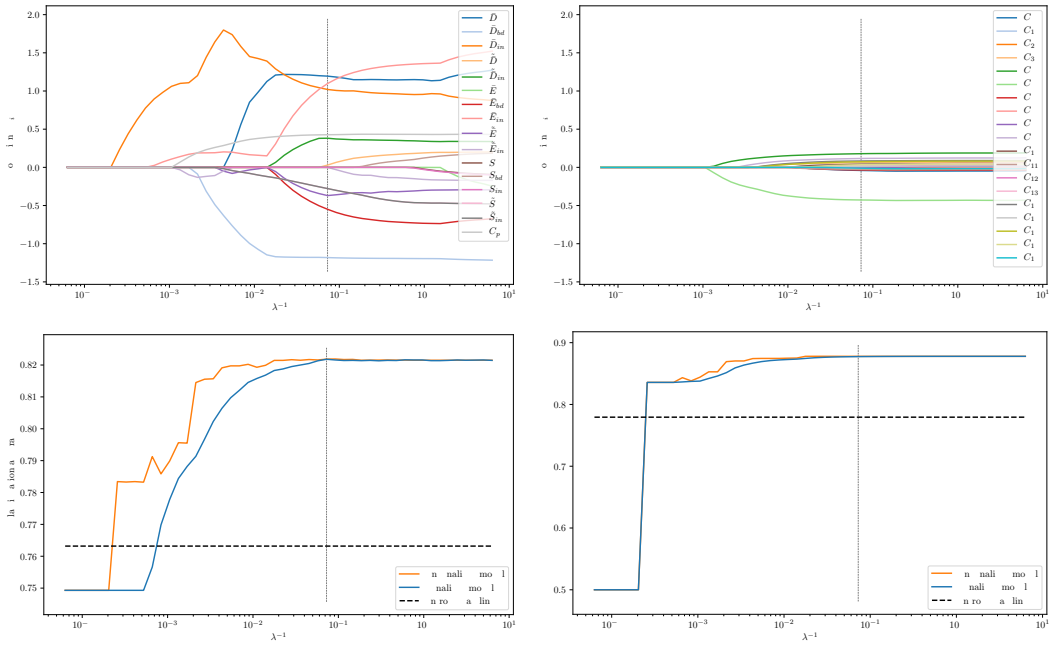


Figure 4: *Results for the meta classification task $IoU_{\text{adj}} = 0, > 0$ for predictions obtained from the Xception65 net. (Top left): the weights coefficients for the 15 metrics computed with LASSO fits as function of λ^{-1}, C_p denotes the maximum of the absolute values of all weight coefficients for predicted classes. (Top right): like top left but showing coefficients for the 18 predicted classes. (Bottom left): meta classification rates for $IoU_{\text{adj}} = 0, > 0$. The blue line are the LASSO fits for different λ values, the orange line shows the performance of regular logistic regression fits ($\lambda = 0$) where the input metrics are only those that have non-zero coefficients in the LASSO fit for the current λ. (Bottom right) same as bottom left, but for AUROC. The vertical dashed lines indicate the λ value for which we obtained the best validation accuracy.*

3 Numerical Experiments: Street Scenes

We investigate the properties of the metrics defined in the previous section for the example of a semantic segmentation of street scenes. To this end, we consider the DeepLabv3+ network [2] for which we use a reference implementation in Tensorflow [1] as well as weights pretrained on the Cityscapes dataset [4] and available on GitHub. The DeepLabv3+ implementation and weights are available for two network backbones: Xception65, which is a modified version of Xception [3] and is a powerful structure intended for server-side deployment, and MobilenetV2 [16], a fast structure designed for mobile devices. Each of these implementations have parameters tuning the segmentation accuracy. We choose the following best (for Xception65) and worst (for MobilenetV2) parameters in order to perform our analysis on two very distinct networks. Note, that the parameter set for the Xception65 setting also includes the evaluation of the input on multiple scales (averaging the results) which increases the accuracy and also leverages classification uncertainty. We refer to [2] for a detailed explanation of the chosen parameters.

- DeepLabv3+Xception65: output stride 8, decoder output stride 4, evaluation on input scales $0.75, 1.00, 1.25$ – $mIoU = 80.42\%$ on the Cityscapes validation set
- DeepLabv3+MobilenetV2: output stride 16, evaluation on input scale 1.00 – $mIoU = 70.71\%$ on the Cityscapes validation set

Example segmentations and heat maps of the two networks are displayed in fig. 2. For both networks, we consider the output probabilities and predictions on the Cityscapes validation set, which consists of 500 street scene images at a resolution of 2048×1024. We compute the 15 constructed metrics as well as IoU_{adj} for each segment in the segmentations of the images. In order to investigate the predictive power of the metrics, we first compute the Pearson correlation $\rho \in [-1, 1]$ between each feature and IoU_{adj}. We report the results of this analysis in table 1 and provide scatter plots of all

features relative to IoU_{adj} in fig. 6. Note, that in all computations, we only consider connected components with non-empty interior.

For both networks IoU_{adj} shows strong correlation with the mean distances $\bar{D}$ and $\bar{D}_{in}$ as well as with the mean entropies $\bar{E}$ and $\bar{E}_{in}$. On the other hand, the relative counterparts are less correlated with IoU_{adj}. The relative segment size $\tilde{S}$ for the DeepLabv3+MobilenetV2 network shows a clear correlation whereas this is not the case for the more powerful DeepLabv3+Xception65 network.

In order to find more indicative measures, we now investigate the predictive power of the metrics when they are combined. For the Xception65 net, we obtain 45194 segments with non-empty interior of which 11331 have $IoU_{\text{adj}} = 0$. For the weaker MobilenetV2 this ratio is $42261/17671$. We would first like to detect segments with $IoU_{\text{adj}} = 0$, i.e., learn the meta classification task of identifying false positive segments based on our 15 metrics and the segment-wise averaged probability distribution vectors. We term these (standardized) inputs x_k for a segment k. Further, let $y_k = ceil(IoU_{\text{adj}}) = \{0 \text{ if } IoU_{\text{adj}} = 0, 1 \text{ if } IoU_{\text{adj}} > 0\}$. The least absolute shrinkage and selection operator (LASSO, [18]) is a popular tool for investigating the predictive power of different combinations of input variables. We compute a series of LASSO fits, i.e., ℓ_1-penalized logistic regression fits

$$\min_w \left[\sum_i -y_i \log(\tau(w^T x_i)) - (1 - y_i)(1 - \log(\tau(w^T x_i))) + \lambda \|w\|_1 \right] , \tag{4}$$

for different regularization parameters λ and standardized inputs (zero mean and unit standard deviation). Here, $\tau(\cdot)$ is the logistic function. Results for the Xception65 net are shown in fig. 4.

The top left and top right panels show, in which order the weight coefficients w for each metric/predicted class become active. At the same time the bottom left and bottom right panels show, which weight coefficient causes which amount of increase in predictive performance in terms of meta-classification rate and AUROC, respectively. The AUROC is obtained by varying the decision threshold of the logistic regression output for deciding whether $IoU = 0$ or $IoU > 0$.

The first non-zero coefficient activates the $\bar{D}_{in}$ metric, which elevates the predictive power above our reference benchmark of choice, the mean entropy per component $\bar{E}$, which we term *entropy baseline*. Another significant gain is achieved when $\bar{D}_{bd}$ and the predicted classes come into play. Noteworthily we obtain a meta-classification validation accuracy of up to $81.91\%(\pm 0.13\%)$ and an AUROC of up to $87.71\%(\pm 0.15\%)$ for Xception65. And also for the weaker MobilenetV2 we obtain $78.93\%(\pm 0.17\%)$ classification accuracy and $86.77\%(\pm 0.17\%)$ AUROC. We randomly choose 10 50/50 training/validation data splits and average the results, the numbers in brackets denote standard deviations of the averages.

Additionally, the bottom line of fig. 4 shows that there is almost no performance loss when only incorporating some of the metrics proposed by the LASSO trajectory. For both networks the classification accuracy corresponds to a logistic regression trained with unbalanced meta-classes $IoU_{\text{adj}} = 0$ and $IoU_{\text{adj}} > 0$, i.e., we did not adjust the class weights. On average (over the 10 training/validation splits) 6851 components with vanishing IoU_{adj} are detected for Xception65 while 4480 remain undetected, for MobilenetV2 this ratio is $14976/2695$. These ratios can be adjusted by varying the probability thresholds for deciding between $IoU_{\text{adj}} = 0$ and $IoU_{\text{adj}} > 0$. For this reason we state results in terms of AUROC which is independent of this threshold.

Ultimately, we want to predict IoU_{adj} values for all connected components and thus model an uncertainty measure. We now resign from regularization and use a linear regression model to predict the IoU_{adj}. Figure 3 depicts the quality of a single linear regression fit for each of the two segmentation networks. For MobilenetV2 we obtain an R^2 value of $81.48\%(\pm 0.23\%)$ and for Xception65 $74.93\%(\pm 0.22\%)$. Figure 5 illustrates the constructed uncertainty measure with two showcases. Averaged results over 10 runs including standard deviations σ and previous meta classification result are summarized in table 2. In all cases, the presented approach clearly outperforms the entropy baseline. The linear regression models do not overfit the data and note-worthily we obtain prediction standard deviations of down to 0.130 and almost no standard deviation for the averages. The classification accuracy and AUROC results are slightly biased towards the validation results as they correspond to the particular λ value that maximizes the validation accuracy. An additional discussion on the difference (also in performance between) IoU_{adj} and IoU can be found in the appendix.

	Xception65		MobilenetV2	
	training	validation	training	validation
	Classification $IoU_{\text{adj}} = 0, > 0$			
ACC, penalized	$81.88\%(\pm 0.13\%)$	$81.91\%(\pm 0.13\%)$	$78.87\%(\pm 0.13\%)$	$78.93\%(\pm 0.17\%)$
ACC, unpenalized	$81.91\%(\pm 0.12\%)$	$81.92\%(\pm 0.12\%)$	$78.84\%(\pm 0.14\%)$	$78.93\%(\pm 0.18\%)$
ACC, entropy baseline	$76.36\%(\pm 0.17\%)$	$76.32\%(\pm 0.17\%)$	$68.33\%(\pm 0.27\%)$	$68.57\%(\pm 0.25\%)$
AUROC, penalized	$87.71\%(\pm 0.14\%)$	$87.71\%(\pm 0.15\%)$	$86.74\%(\pm 0.18\%)$	$86.77\%(\pm 0.17\%)$
AUROC, unpenalized	$87.72\%(\pm 0.14\%)$	$87.72\%(\pm 0.15\%)$	$86.74\%(\pm 0.18\%)$	$86.76\%(\pm 0.18\%)$
AUROC, entropy baseline	$77.81\%(\pm 0.16\%)$	$77.94\%(\pm 0.15\%)$	$76.63\%(\pm 0.24\%)$	$76.74\%(\pm 0.24\%)$
	Regression IoU_{adj}			
σ, all metrics	$0.181(\pm 0.001)$	$0.182(\pm 0.001)$	$0.130(\pm 0.001)$	$0.130(\pm 0.001)$
σ, entropy baseline	$0.258(\pm 0.001)$	$0.259(\pm 0.001)$	$0.215(\pm 0.001)$	$0.215(\pm 0.001)$
R^2, all metrics	$75.06\%(\pm 0.22\%)$	$74.97\%(\pm 0.22\%)$	$81.50\%(\pm 0.23\%)$	$81.48\%(\pm 0.23\%)$
R^2, entropy baseline	$49.37\%(\pm 0.32\%)$	$49.02\%(\pm 0.32\%)$	$49.32\%(\pm 0.31\%)$	$49.12\%(\pm 0.32\%)$

Table 2: *Summarized results for classification and regression, averaged over 10 runs. The numbers in brackets denote standard deviations of the computed mean values.*

4 Conclusion and Outlook

We have shown statistically that per-segment metrics derived from entropy, probability difference, segment size and the predicted class clearly contain information about the reliability of the segments and constructed an approach for detecting unreliable segments in the network's prediction. In our tests with publicly available pre-trained DeepLabv3+ networks the computed logistic LASSO fits for meta classification task $IoU_{\text{adj}} = 0$ vs. $IoU_{\text{adj}} > 0$ obtain AUROC values of up to 87.71% and classification rates of up to 81.91%. When predicting the IoU_{adj} with a linear regression fit we obtain a prediction standard deviation of down to 0.130, as well as R^2 values of up to 81.48%. These results could be further improved when incorporating model uncertainty in heat map generation. We believe that using MC dropout will further improve these results, just like the development of ever more accurate networks. We plan to use our method for detecting labeling errors, for label acquisition in active learning and we plan to investigate further metrics that may leverage detection accuracy. Apart from that, detection mechanisms built on the softmax input and even earlier layers could be thought of. The source code of our method is publicly available at `https://github.com/mrottmann/MetaSeg`.

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

References

[1] M. Abadi, A. Agarwal, P. Barham, et al. *TensorFlow: large-scale machine learning on heterogeneous systems*. Software available from tensorflow.org. 2015.

[2] L.-C. Chen et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation". *Corr* abs/1802.02611 (2018).

[3] F. Chollet. "Xception: deep learning with depthwise separable convolutions". *2017 ieee conference on computer vision and pattern recognition (cvpr)* (2017), pp. 1800–1807.

[4] M. Cordts et al. "The cityscapes dataset for semantic urban scene understanding ". *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[5] F. Cowell. *Measuring Inequality*. Oxford Univ. Press, 2011.

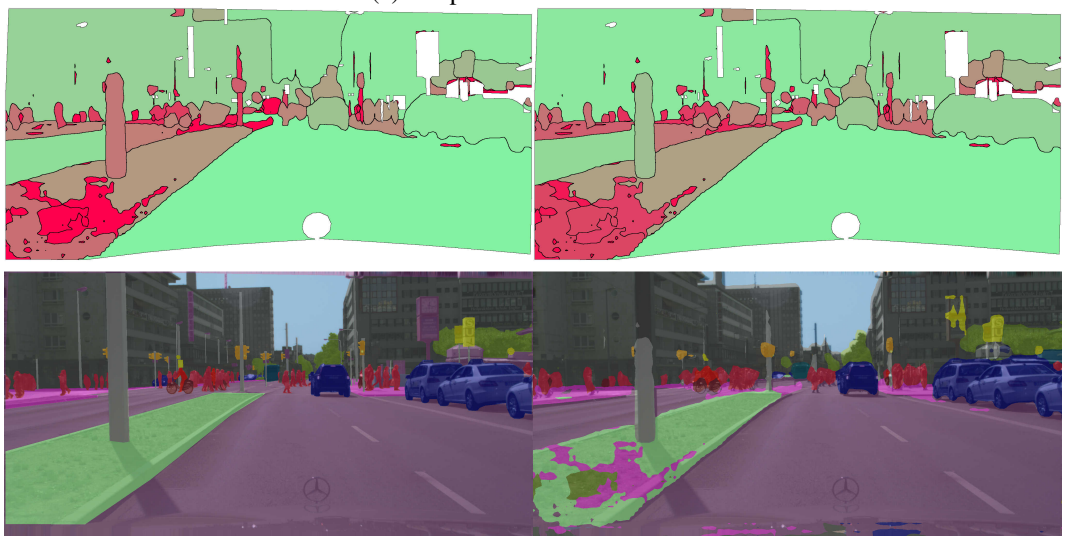
[6] J. Davis and M. Goadrich. "The relationship between precision-recall and ROC curves ". *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*. 2006, pp. 233–240.

[7] Y. Gal and Z. Ghahramani. "Dropout as a bayesian approximation: representing model uncertainty in deep learning ". *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.

[8] D. Hendrycks and K. Gimpel. "A baseline for detecting misclassified and out-of-distribution examples in neural networks". *Corr* abs/1610.02136 (2016).

[9] P.-Y. Huang et al. "Efficient uncertainty estimation for semantic segmentation in videos ". *European Conference on Computer Vision (ECCV)*. 2018.

(a) DeepLabv3+ MobilenetV2



(b) DeepLabv3+ Xception65

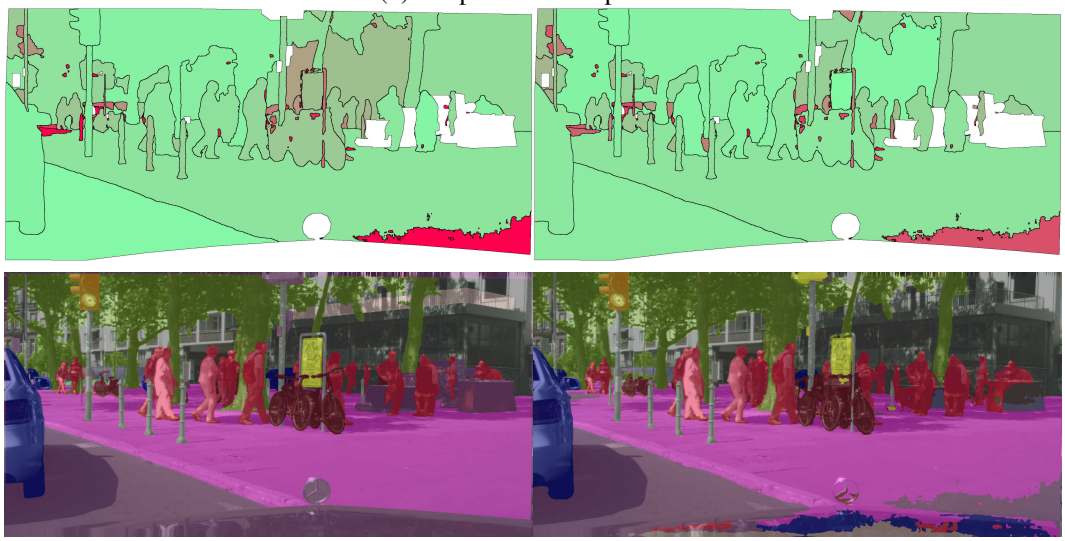


Figure 5: *Prediction of the IoU_{adj} with linear regression. Each of the two sub-figures (a) and (b) consists of ground truth (bottom left), predicted segments (bottom right), true IoU_{adj} for the predicted segments (top left) and predicted IoU_{adj} for the predicted segments (top right). In the top row, green color corresponds to high IoU_{adj} values and red color to low ones, for the white regions there is no ground truth available. These regions are excluded from the statistical evaluation.*

[10] P. Jaccard. "The distribution of the flora in the alpine zone". *New phytologist* 11.2 (Feb. 1912), pp. 37–50.

[11] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen. "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks". *2016 ieee conference on computer vision and pattern recognition workshops (cvprw)* (2016), pp. 680–688.

[12] A. Kendall, V. Badrinarayanan, and R. Cipolla. "Bayesian segnet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding". *Corr* abs/1511.02680 (2015).

[13] S. Liang, Y. Li, and R. Srikant. "Principled detection of out-of-distribution examples in neural networks". *Corr* abs/1706.02690 (2017).

[14] W. Lin and A. Hauptmann. "Meta-classification: combining multimodal classifiers". *Mining multimedia and complex data. pakdd 2002. lecture notes in computer science* 2797 (2003).

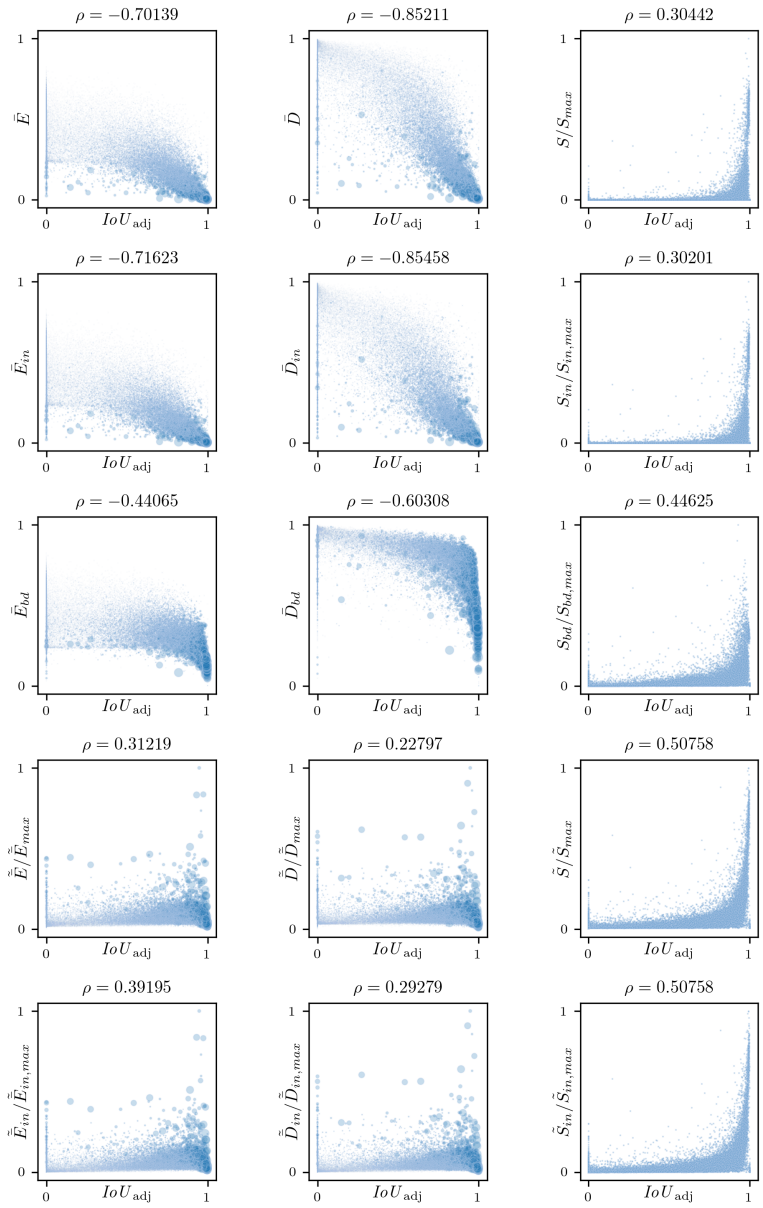


Figure 6: *Correlations between IoU_{adj} and rescaled features for the DeepLabv3+Xception65 network. Dot sizes in the first two columns are proportional to S.*

[15] P. Oberdiek, M. Rottmann, and H. Gottschalk. "Classification uncertainty of deep neural networks based on gradient information ". *Artificial Neural networks and Pattern Recognition (ANNPR)*. 2018.

[16] M. Sandler et al. "Inverted residuals and linear bottlenecks: mobile networks for classification, detection and segmentation". *Corr* abs/1801.04381 (2018).

[17] C. E. Shannon. "A mathematical theory of communication ". In: *The Bell System Technical Journal*. Vol. 27. 1948, pp. 379–423, 623–656.

[18] R. Tibshirani. "Regression shrinkage and selection via the lasso". *Journal of the royal statistical society: series b* 58 (1996), pp. 267–288.

[19] K. Wickstrøm, M. Kampffmeyer, and R. Jenssen. "Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps". *Corr* abs/1807.10584 (2018).